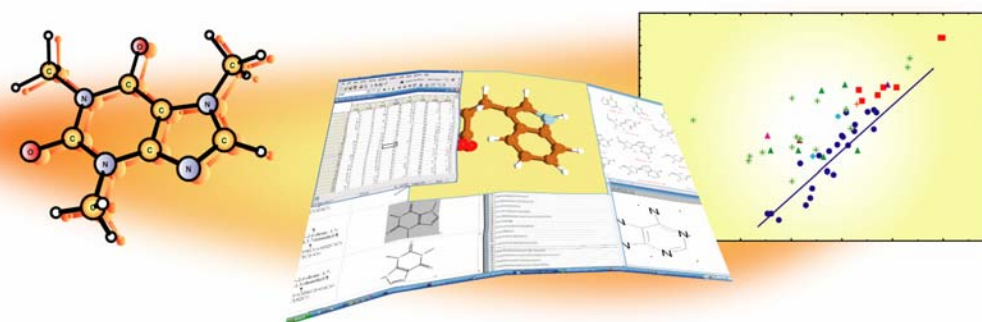




Collection and Evaluation of (Q)SAR Models for Mutagenicity and Carcinogenicity

Romualdo Benigni, Cecilia Bossa, Tatiana Netzeva, Andrew Worth





EUROPEAN COMMISSION
DIRECTORATE GENERAL
JOINT RESEARCH CENTRE

Institute for Health and Consumer Protection
Toxicology and Chemical Substances Unit
European Chemicals Bureau
I-21020 Ispra (VA) Italy

Collection and Evaluation of (Q)SAR Models for Mutagenicity and Carcinogenicity

Romualdo Benigni¹, Cecilia Bossa¹, Tatiana Netzeva², Andrew Worth²

¹Health and Environment Department, Istituto Superiore di Sanita, Rome, Italy

²European Chemicals Bureau, Institute for Health and Consumer Protection,
Joint Research Centre, European Commission, Ispra (VA), Italy

2007

EUR 22772EN



EUROPEAN COMMISSION
DIRECTORATE-GENERAL
Joint Research Centre



The mission of IHCP is to provide scientific support to the development and implementation of EU policies related to health and consumer protection.

The IHCP carries out research to improve the understanding of potential health risks posed by chemicals, physical and biological agents from various sources to which consumers are exposed.

- As a Research Institute, the IHCP contributes to the improvement of scientific knowledge on health care methods and consumer issues.
- As a European Institution, the IHCP participates, at an international level and in collaboration with Member States Authorities, in R&D and regulatory actions intended to improve consumer tutelage.
- As a European Commission Service, the IHCP acts as an independent advisor for the implementation of risk assessment, monitoring and validation of procedures aimed to insure EU citizens of the use of health and consumer services or products.

European Commission
Directorate-General Joint Research Centre
Institute IHCP

Contact information Andrew Worth
Address: European Chemicals Bureau, TP81
E-mail: Andrew.Worth@jrc.it
Tel.: +39 0332 789566
Fax: +39 0332 786717

<http://ihcp.jrc.cec.eu.int/>; <http://ecb.jrc.it/>
<http://www.jrc.cec.eu.int>

Legal Notice

Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use which might be made of this publication.

A great deal of additional information on the European Union is available on the Internet.
It can be accessed through the Europa server
<http://europa.eu.int>

22772 EN

ISSN 1018-5593

Luxembourg: Office for Official Publications of the European Communities

© European Communities, 2007

Reproduction is authorised provided the source is acknowledged

Printed in Italy

Summary

This evaluation of the non-commercial (Q)SARs for mutagenicity and carcinogenicity consisted of a preliminary survey (Phase I), and then of a more detailed analysis of short listed models (Phase II). In Phase I, the models were collected from the literature, and then assessed according to the OECD principles –based on the information provided by the authors-. Phase I provided the support for short listing a number of promising models, that were analyzed more in depth in Phase II. In Phase II, the information provided by the authors was completed and complemented with a series of analyses aimed at generating an overall profile of each of the short listed models.

The models can be divided into two families based on their target: a) congeneric; and b) non-congeneric sets of chemicals.

The QSARs for congeneric chemicals include most of the chemical classes top ranking in the EU High Production Volume list, with the notable exception of the halogenated aliphatics. They almost exclusively aim at modeling Salmonella mutagenicity and rodent carcinogenicity, which are crucial toxicological endpoints in the regulatory context. The lack of models for *in vivo* genotoxicity should be remarked. Overall the short listed models can be interpreted mechanistically, and agree with, and/or support the available scientific knowledge, and most of the models have good statistics. Based on external prediction tests, the QSARs for the potency of congeneric chemicals are 30 to 70 % correct, whereas the models for discriminating between active and inactive chemicals have considerably higher accuracy (63 to 100 %), thus indicating that predicting intervals is more reliable than predicting individual data points. The internal validation procedures (e.g., cross-validation, etc...) did not seem to be a reliable measure of external predictivity.

Among the non-local, or global approaches for non-congeneric data sets, four models based on the use of Structural Alerts (SA) were short listed and investigated in more depth. The four sets did not differ to a large extent in their performance. In the “general” databases of chemicals the SAs appear to agree around 65% with rodent carcinogenicity data, and 75% with Salmonella mutagenicity data.

The SAs based models do not seem to work equally efficiently in the discrimination between active and inactive chemicals within individual chemical classes. Thus, their main role is that of preliminary, or large-scale screenings. A priority for future research on the SAs is their expansion to include alerts for nongenotoxic carcinogens.

A general indication of this study, valid for both congeneric and noncongeneric models, is that there is uncertainty associated with (Q)SARs; the level of uncertainty has to be considered when using (Q)SAR in a regulatory context. However, (Q)SARs are not meant to be black-box machines for predictions, but have a much larger scope including organization and rationalization of data, contribution to highlight mechanisms of action, complementation of other data from different sources (e.g., experiments). Using only non-testing methods, the larger the evidence from QSARs (several different models, if available) and other approaches (e.g. chemical categories, read across) the higher the confidence in the prediction.

Content

1	Introduction.....	5
2	The QSARs for congeneric classes of chemicals.....	12
2.1	General considerations	12
2.2	Short listed QSARs	15
2.3	Approach to the characterization of the short listed models	16
2.4	Characterization of the individual QSARs: results of the survey	19
2.4.1	Aromatic amines	19
2.4.1.1	QSAR 1, Ref. 4 in Appendix 1, Debnath et al., 1992, mutagenic potency in <i>Salmonella typhimurium</i> TA98.....	20
2.4.1.2	QSAR 2, Ref. 4 in Appendix 1, Debnath et al., 1992, mutagenic potency in <i>Salmonella typhimurium</i> TA100.....	24
2.4.1.3	QSAR 3, Ref. 23 in Appendix 1, Benigni et al., 2000, carcinogenic potency in mouse	28
2.4.1.4	QSAR 4, Ref. 23 in Appendix 1, Benigni et al., 2000, carcinogenic potency in rat	32
2.4.1.5	QSAR 5, our unpublished results, mutagenic activity in <i>Salmonella typhimurium</i> TA98	35
2.4.1.6	QSAR 6, our unpublished results, mutagenic activity in <i>Salmonella typhimurium</i> TA100	39
2.4.1.7	QSAR 7, Ref. 24 in Appendix 1 (Eq. 4 in the paper), Franke et al., 2001, carcinogenic activity in rodents (overall).....	43
2.4.1.8	QSAR 8, Ref. 24 (Eq. 5 in the paper) in Appendix 1, Franke et al., 2001, carcinogenic activity in rodents (overall).....	47
2.4.2	Nitroarenes	52
2.4.2.1	QSAR 9, Ref. 32, mutagenic potency in <i>Salmonella typhimurium</i> TA98	52
2.4.2.2	QSAR 10, Ref. 33, mutagenic potency in <i>Salmonella typhimurium</i> TA100	56
2.4.3	Polycyclic aromatic hydrocarbons (PAH)	59
2.4.3.1	QSAR 11, Ref. 47 in Appendix 1, Zhang et al., 1992, Skin carcinogenicity in rodents	59
2.4.4	α,β -Unsaturated aliphatic aldehydes	63
2.4.4.1	QSAR 12, Ref. 60 in Appendix 1, Benigni et al., 2003, mutagenic potency in <i>Salmonella typhimurium</i> TA100.....	63
2.4.4.2	QSAR 13, Refs. 60 and 61 in Appendix 1, Benigni et al., 2003; 2005, mutagenic activity in <i>Salmonella typhimurium</i> TA100.....	66
3	Non-local (Q)SARs.....	69
3.1	The short listed models	69
3.2	Building the capacity to manipulate the SAs	70
3.3	Probes for the SAs: the databases	71
3.4	The characterization of the SAs-based models.....	72
3.4.1	Ashby' SAs	72
3.4.2	Bailey' SAs	76
3.4.3	Kazius et al., 2005, SAs	78
3.4.4	Kazius et al., 2006, SAs	81
3.4.5	Non-general databases as probes	82
4	Conclusions	86
4.1	QSARs for congeneric sets of chemicals	86
4.2	SARs for non-congeneric data sets	92
4.3	Final considerations	95

5	References.....	98
6	Scheme 1: OECD PRINCIPLES-BASED SCORING SYSTEM FOR THE (Q)SAR MODELS 102	
7	Appendix 1 List of reviewed papers	104
8	Appendix 2 Scoring results for selected models.....	111
9	Appendix 3 Regression-based models for mutagenicity and carcinogenicity	113

1 Introduction

The work performed for this project by the Istituto Superiore di Sanita' is part of a wide range of initiatives coordinated by the European Chemicals Bureau – Joint Research Centre for the technical preparations necessary to implement REACH in EU. ECB coordinates the JRC Action on Computational Toxicology, which aims to promote the availability of valid (Q)SARs and related estimation methods for possible regulatory use. This case study on the validation of non-commercial (Q)SAR models for mutagenicity and carcinogenicity follows previous case studies coordinated by ECB on selected endpoints, such as acute fish toxicity, skin sensitization, skin penetration, binding to the estrogen and androgen receptors. This project considers only the non-commercial (Q)SAR models for mutagenicity and carcinogenicity. For this reason, very popular models obtained through the application of e.g., MultiCase, Topkat, DEREK, ADAPT and PASS software systems were excluded from this study.

According to the lines indicated in the contract, the first step of the work was a survey of the literature, and the collection of the existing (Q)SAR models. This was followed by an evaluation organized in two phases. In Phase 1, the models were evaluated based on the information reported by the authors. A scoring system inspired to the OECD principles was designed by us to provide support in this task. For the QSAR models focusing on individual chemical classes, additional criteria were the industrial / environmental importance of the chemical class. The scores from Phase 1 pointed to a short list of most promising models, to be considered in depth in Phase 2. Among others, the transparency and availability of information on the models has played a crucial role in Phase 1. In Phase 2, for each of the short-listed models extensive validation work was carried out, consisting of: Level A) verification of the algorithm and relative statistics, cross-validation, exploration of the applicability domain; Level B) whenever feasible, validation using an independent test set (assessment of predictivity).

This work generated assessments of the short listed models, as well as overall final considerations.

To present the results of Phase I, and to put them in a wider perspective, a Workshop on (Q)SAR Modeling of Mutagenicity and Carcinogenicity was co-organised by the Istituto Superiore di Sanita and ECB (Rome, 22 - 23 June 2006; Romualdo Benigni, Chairman), with contributions provided by a number of researchers from academia, regulatory authorities and private companies. The presentations at the Workshop were summarized in a report paper, recently submitted for publication. The paper provides a review and state of the art in the modeling of mutagenicity and carcinogenicity as well as views and opinions of the individual authors on more general issues such as validity of (Q)SAR models and reliability of their predictions.

Collection, filtering and scoring of (Q)SAR models

A total of 78 non-commercial (Q)SAR models were collected from the literature (see list in Appendix 1). The list categorizes the models into the classes of: 1) models for congeneric sets of chemicals. These models were mainly obtained through the application of the Hansch approach, or of Hansch-like discriminant analysis; 2) non-local, or general models for noncongeneric sets of chemicals. Group 1 is further subdivided according to the chemical class and the toxicological end-point.

The models in Appendix 1 vary to a large extent in terms of scientific value, breadth and scope. Preliminary to short-listing the models relevant to the EU regulatory needs, first a “quick” filter was applied, followed by a more systematic scoring of the “surviving” models. It should be emphasized that all the listed models contribute with interesting pieces of information. Even if not relevant immediately to the regulatory scopes, they may be the basis for further future developments; thus, it is important to keep track of them.

The “quick” filter eliminated the following (classes of) models (see Appendix 1):

- a) non regulatory end-points, outside REACH scopes (e.g., models for metabolism, for pharmaceuticals and food, non official methods): 15, 27, 43, 44, 45, 51, 52, 53, 54, 67, 78;
- b) methodological studies e.g., focusing on mechanism representation, or parameters validation: 2, 3, 16, 22, 37, 38, 39, 43 ;
- c) models replaced by subsequent, more refined models: 1, 17, 18, 19, 20, 29, 30, 31, 51, 66, 75;
- d) models based on very small data sets ($n < 15$): 40, 41, 55, 57, 58, 59.

In total, 36 papers were excluded from a more formal evaluation.

The models that survived the above step were classified and given scores, by and large based on the OECD principles. The scoring system was elaborated by us specifically for this work: It is

provided in Scheme 1. The scoring system is aimed at providing characterization profiles of the individual models, as a basis for the subsequent short listing. For each heading, the best score is 1; higher figures correspond to lower quality ratings. In this way, the best profile is that of models with the lowest overall numerical figures.

It should be emphasized that the score system adopted in this phase characterizes the models, based on the information provided by the authors in the original papers. Thus, the scores are only weakly related to the scientific value of the models and mostly reflect the availability of information necessary for the short-listing of promising models, and their subsequent in-depth evaluation.

In Scheme 1: Point A regards the clarity with which the algorithm is described; Point B regards the applicability domain; , Point C regards the statistical documentation; Point D regards the interpretability in mechanistic terms. In addition to the OECD principles, we considered in our scoring system also: E1) the number of data points used by the authors; and E2) the number of chemicals of the same class present in the HPV inventory.

The relevance of E1 is self-evident. Regarding E2 it should be noted that, to assess the regulatory relevance of the (Q)SAR models, it is also important to weight the models against the pattern of use of chemicals in the real life. A realistic parameter is the presence of the various chemical classes in the European list of High Production Volume (HPV) chemicals. To this aim, we downloaded the list of EU HPV from the ESIS website, and we constructed the database of structures. This database of HPV structures was checked for the distribution of the chemical classes for which QSARs exist in the literature, and the results of this interrogation are the E2 scores.

The scores of the models are in Appendix 2. The identification number of each model is in Appendix 1; the repartition in chemical classes is reported in Appendix 1 as well. As additional information, Appendix 3 reports some documentation on the QSARs considered.

The models differed to a large extent in their characteristics. Figure 1 shows the distribution of the average scores. Figures 2 and 3 show the distribution of data points used by the authors to generate the QSARs for the individual chemical classes, and the non-local (Q)SARs, respectively. Thus the

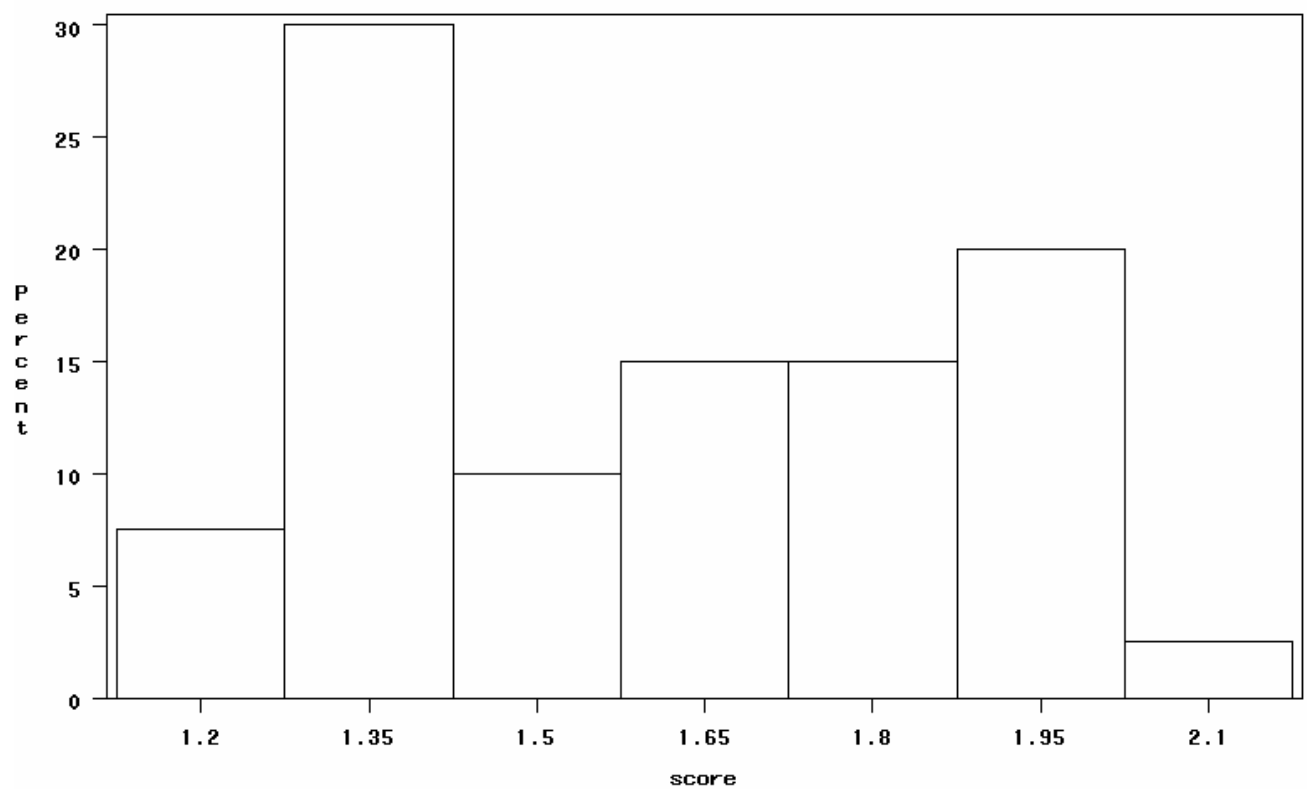


Figure 1

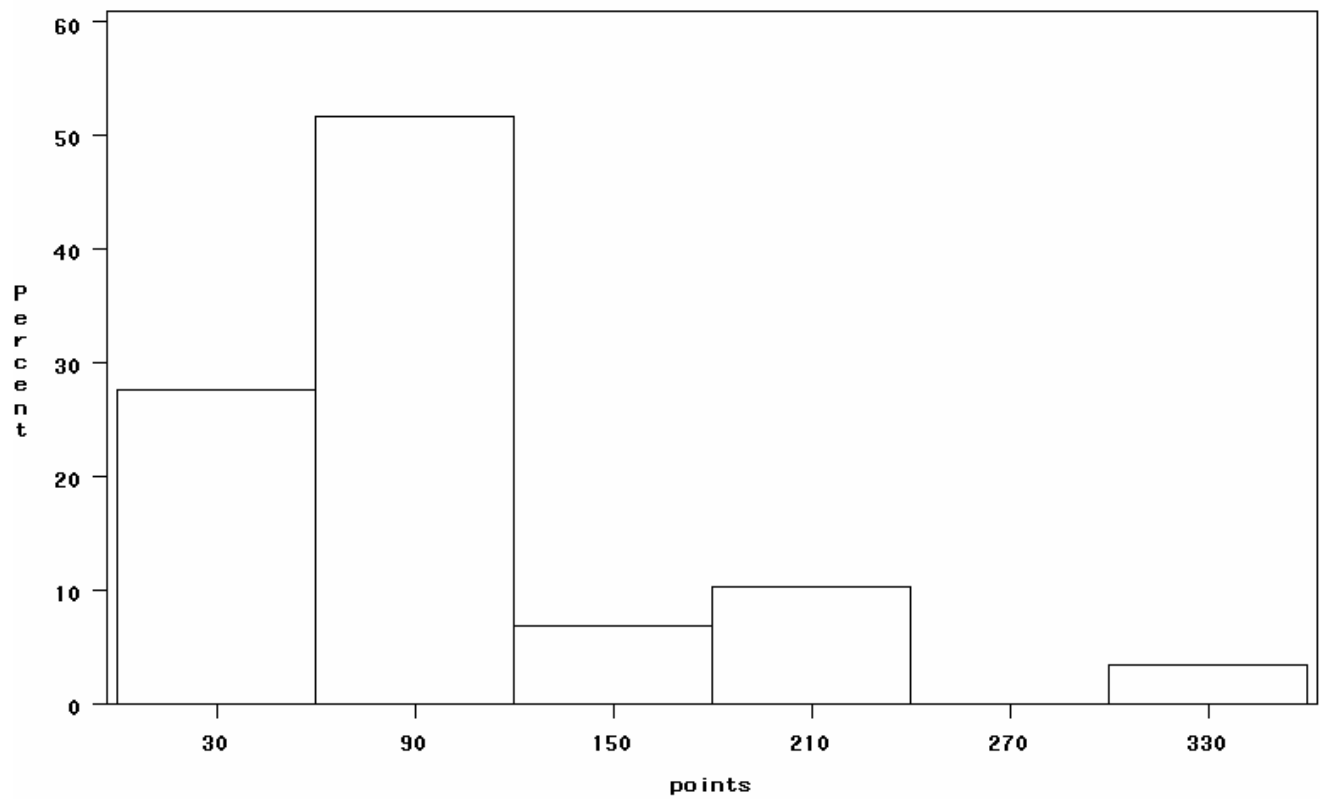


Figure 2

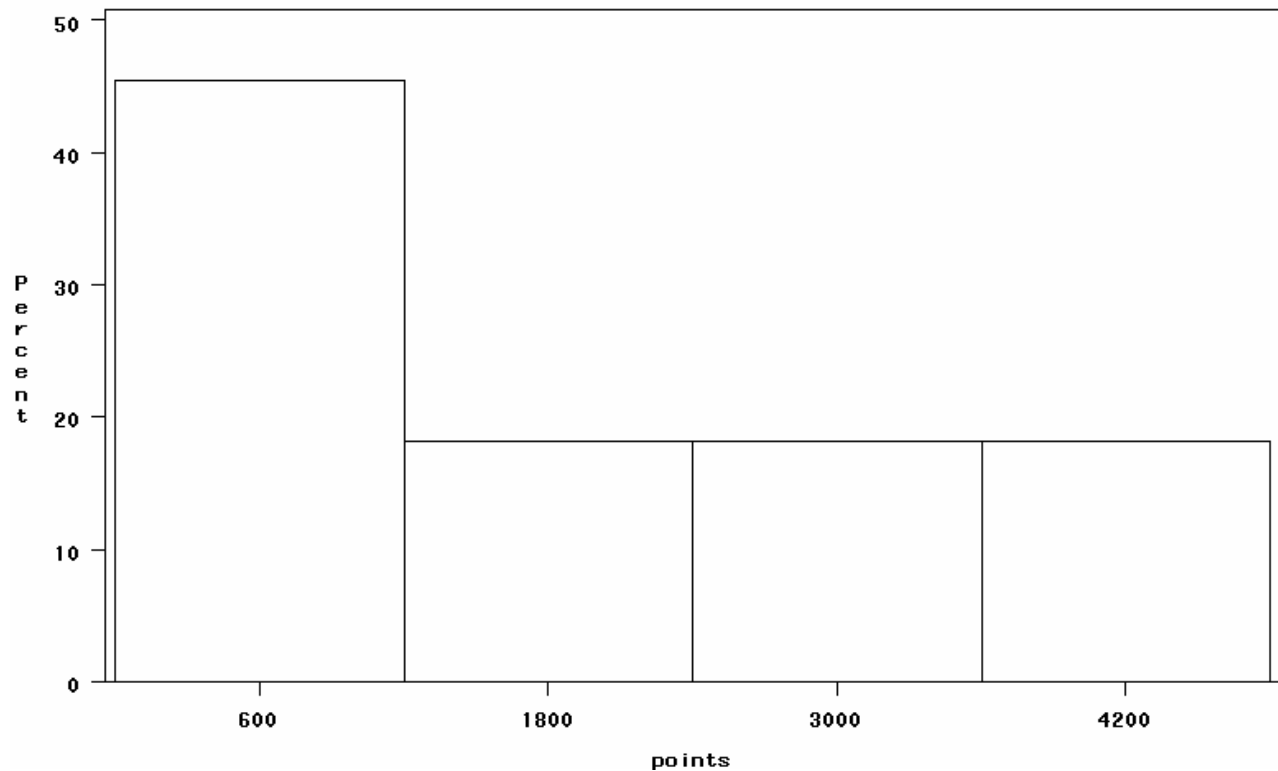


Figure 3

documentation in Appendix 2 provided a useful framework to short list a selection of promising models, that were characterized more in depth. It should be emphasized that the documentation in Appendix 2 was used in a flexible way. For example some models (e.g., Ref. 62, Ashby' Alerts) cannot be given scores, since it was not derived in a formalized way and many criteria cannot be applied (the model comes from the non-formal accumulation of various sources of information on the mechanisms of action). However, for their scientific importance and impact on research such models cannot be excluded from consideration.

The flexible application of our subjective judgment to Appendix 2 produced the list of promising models reported in Table I. The numbers under Reference identify the models in Appendix 1. Detailed discussion and analysis of the individual models is in the following sections.

Table I: Short listed (Q)SAR models

<i>Code</i>	<i>Chemical class</i>	<i>Biological endpoints</i>	<i>Type</i>	<i>Reference</i>
QSAR1	Aromatic amines	Salmonella mutagenicity TA98	Potency	4
QSAR2	“ “	“ “ TA100	“	4
QSAR3	“ “	Mouse Carcinogenicity	“	23
QSAR4	“ “	Rat Carcinogenicity	“	23
QSAR5	“ “	Salmonella mutagenicity TA98	Activity	Our Results
QSAR6	“ “	“ “ TA100	“	“ “
QSAR7	“ “	Rodent overall Carcinogenicity	“	24
QSAR8	“ “	Rodent overall Carcinogenicity	“	24
QSAR9	Nitroarenes	Salmonella mutagenicity TA98	Potency	32
QSAR10	“ “	“ “ TA100	“	33
QSAR11	PAH	Rodent skin Carcinogenicity	“	47
QSAR12	Aliphatic Aldehydes	Salmonella mutagenicity TA100	Potency	60
QSAR13	“ “	“ “ “	Activity	60,61
Ashby’ SAs	All	Genotoxic Carcinogenicity	“	62
Bailey’s SAs	“	“ “	“	65
Kazius’ SAs (1)	“	Salmonella Mutagenicity	“	63
Kazius’ SAs (2)	“	“ “	“	64

nitroaromatic compounds and aliphatic aldehydes. On the contrary, the polycyclic aromatic hydrocarbons, for which a number of QSARs exist, are a minority as pure chemicals in the EU HPV inventory. However, several of them are present as components of petrol products.

An additional information obtained by analysing the HPV database is that about one sixth of the structurally coded chemicals (~16%) on the EU HPV list triggers at least one SA (as given in Ashby's compilation). This gives a rough estimation of the carcinogenic risk posed by the HPV chemicals, and of the extent of work necessary to evaluate this finding further.

Table II: QSARs and EU HPV chemicals

Classes with QSARs	Number of HPV chemicals
Aromatic amines	115
Halogenated aliphatics	113
Nitroaromatic compounds	42
Aliphatic aldehydes	33
Epoxides	10
Polycyclic Aromatic Hydrocarbons	6
Lactones	3
Quinolines	1

2.2 Short listed QSARs

Among the QSAR models for individual chemical classes, the models selected in Phase 1 of our survey are relative to the classes of aromatic amines, nitroarenes, aliphatic aldehydes, and polycyclic aromatic hydrocarbons (PAH). As shown in Table II, the first 3 classes are those most represented among the HPV chemicals; on the other hand, the PAHs are poorly represented as pure and well defined chemicals, but are widespread as mixtures or impurities. Halogenated aliphatics are one of top ranking classes in Table II; however QSARs exist only for one endpoint (Aneuploidy in *Aspergillus nidulans*) that is valid scientifically, but is not included among the recognized regulatory methods (Refs. 51, 52, 53 in Appendix 1).

The models selected are described in the following references of Appendix 1: a) Ref 4 (mutagenicity) and Refs. 23 and 24 (carcinogenicity) for aromatic amines; b) Refs. 32 and 33 (mutagenicity) for nitroarenes; c) Refs. 60 and 61 (mutagenicity) for aliphatic aldehydes; d) Ref. 47 (carcinogenicity) for PAHs.

Appendix 2 shows that all these models: a) have high ratings for Points A1, A2, A3 of Scheme 1 (they provide a documentation sufficient for the models to be judged and analyzed in depth); b) have high ratings for Point D as well (they can be interpreted in mechanistic terms); and c) they rely on a sufficient number of data points, and they refer to some of the most represented classes among the HPV chemicals (Table II).

The models are widely different for Points B and C of Scheme 1 (for example, some have information on external predictivity and some do not), but the information provided by the authors allowed us to re-check the results of the authors, and to complete the documentation.

2.3 Approach to the characterization of the short listed models

The characterization of the individual, short listed models follows the general scheme of OECD principles.

A defined endpoint

The association of the model with a defined toxicity endpoint, addressed by an officially recognised test method (Annex V to *Directive 67/548/EEC*), is reported.

An unambiguous algorithm

The model given by the authors (together with the associated statistics) is reported.

A defined domain of applicability

The applicability domain of the model is defined in terms of: a) the structures to which it applies; and b) the range of the values of the descriptors in the model.

If this information is not reported by the authors, it is derived by us.

Appropriate measures of goodness-of-fit, robustness and predictivity

The model is re-analyzed by us, and the concordance with the measures reported by the authors is given. When necessary, additional measures are calculated by us.

For Multiple Linear Regression (MLR) models, the final overall characterization includes: r^2 ; Adjusted r^2 ; q^2 .

The Cross-Validated r^2 (q^2) is $= 1 - (\text{sum of squares of the predictive residuals} / \text{sum of squares of the mean-centered response data})$.

The mean Leverage (with SD) is used to assess if a model depends in a balanced way on all the data points (corresponding to low mean Leverage).

For discriminant models (Linear Discriminant Analysis and Canonical Discriminant Analysis), the final overall characterization includes: accuracy, sensitivity, and specificity, together with the Squared Canonical Correlation.

Accuracy is the percentage of all chemicals correctly identified by the model. Sensitivity is the percentage of biologically active (positive) chemicals correctly identified (calculated out of the total number of positives). Specificity is the percentage of biologically inactive (negative) chemicals correctly identified (calculated out of the total number of negatives).

The Squared Canonical Correlation is a measure of the correlation between the biological activity variable, and the linear combination of descriptor variables (produced either by Linear Discriminant Analysis or Canonical Discriminant Analysis) that best separates the negatives from the positives.

For all models, the results of cross-validation is given. Three leave-many-out procedures were considered, leaving out: a) 10%; b) 25%; and c) 50% of the data set. Each procedure was applied ten times (by random selection of excluded chemicals; see more details in the following sections). For the MLR models, the average q^2 (with Standard Deviation (SD)) is reported. For discriminant models, accuracy, sensitivity, and specificity (with SD) are reported.

If the QSAR model is assessed by the authors for its predictivity of external compounds, the results are given (together with the results of our re-checking). Otherwise, external data sets were sought by us, and (when available) the details on the external prediction exercise are reported.

For regression based models, the performance in external validation is expressed as correlation coefficient between experimental and predicted potency. An additional way of measuring the

prediction performance is the percentage of test chemicals correctly predicted within one log unit of potency.

For discriminant models, the external predictivity is measured as percentage of test chemicals correctly predicted (accuracy).

Together with performance measures, the degree of concordance between the chemical domains of the training and test sets is reported in terms of: a) types of chemical structures; b) range of the values of the descriptors; c) chemical similarity indices.

For Point c), the training and test sets are combined and the overall Tanimoto similarity matrix is calculated with the computer software Leadscape. To smooth the effect of “weird” individual similarities and to filter the data by exploiting the global relationships patterns, an Euclidian distance matrix is calculated from the similarity matrix, and then subjected to Principal Component Analysis (Sneath 1983; Sneath & Johnson 1972) (Benigni 1993). Finally, the ranges of PC scores for the training and test sets are compared.

A mechanistic interpretation, if possible

The concordance of the model with mechanistic knowledge is discussed.

Abbreviations of general significance

HOMO is the energy of the Highest Occupied Molecular Orbital,

LUMO is the energy of the Lowest Unoccupied Molecular Orbital.

2.4 Characterization of the individual QSARs: results of the survey

2.4.1 Aromatic amines

The aromatic amines are one of the chemical classes with the largest environmental and industrial impact, and are the most represented class among the HPV chemicals (Table II). Consequently, this class has been the subject of a large number of QSAR analyses (Point 1.1 in Appendix 1), focusing both on their mutagenic (mainly) and carcinogenic effects.

The short listed QSARs are (Appendix 1):

Ref. 4, mutagenic potency in *Salmonella typhimurium* TA98 and TA100;

Ref.23, carcinogenic potency in rodents;

Ref. 24, carcinogenic activity in rodents (two alternative models).

It should be noted that activity refers to the difference between active and inactive compounds, whereas potency refers to the gradation of the biological effect among the active chemicals only. As a matter of fact, for this chemical class the QSAR models for the mutagenic potency are different from those for the activity (Benigni, Andreoli & Giuliani 1994; Franke, Gruska, Giuliani & Benigni 2001).

No models for the mutagenic activity of the aromatic amines exist in the literature, so a QSAR analysis was specifically performed by us for this project, based on the data in Ref. 4.

Refs. 4, 23, and 24 report some statistical measures of fitting and internal validation, however no check of external predictivity are reported. Thus, in addition to internal validation checks (e.g., LMO, etc...) we found in the literature further chemicals suitable for external validation. For each model, the results of our analyses are reported here.

2.4.1.1 QSAR 1, Ref. 4 in Appendix 1, Debnath et al., 1992, mutagenic potency in *Salmonella typhimurium* TA98

A defined endpoint

This QSAR is associated with a defined toxicity endpoint (genetic mutation), addressed by an officially recognised test method (Method B.13/14 Mutagenicity – Reverse Mutation test using bacteria – Annex V to *Directive 67/548/EEC*).

An unambiguous algorithm

The data and QSAR models are reported in Ref. 4. The mutagenic potency in TA98 strain (+ S9 activation system) was modelled by:

$$\log \text{TA98} = 1.08(\pm 0.26) \log P + 1.28(\pm 0.64)\text{HOMO} - 0.73(\pm 0.41)\text{LUMO} + 1.46(\pm 0.56)IL + 7.20(\pm 5.4) \quad (4.6)$$

$$n = 88, r = 0.898 \ (r^2 = 0.806), s = 0.860, F_{1,83} = 12.6$$

The mutagenic potency (log TA98) is expressed as log (revertants/nmol). High TA98 values indicate high mutagenic potency. The AM1 molecular orbital energies (HOMO and LUMO) are given in eV. *IL* is an indicator variable that assumes a value of 1 for compounds with three or more fused rings.

A defined domain of applicability

The applicability domain of the model is defined explicitly by the authors in terms of the structures to which it applies. The chemical set spans a large range of basic structures (aniline, biphenyl, anthracene, phenanthrene, fluorene, pyrene, fluoranthene, chrysene, quinoline, carbazole, phenazine) with one or two amine groups attached.

A number of outliers were excluded from the analysis, but they do not point to general rules for defining the applicability domain.

The ranges of the chemical descriptors are not reported explicitly, and were derived by us:

logP: 1.12 4.98;
HOMO: -10.018 -7.528;
LUMO: -1.691 0.722;

Appropriate measures of goodness-of-fit, robustness and predictivity

The goodness-of-fit reported by the authors is the correlation coefficient (see above). The data were re-analysed by us for this work. A Multiple Linear Regression (MLR) analysis reproduced the original QSAR equation, with

$r^2 = 0.807$;

Adjusted $r^2 = 0.798$;

$q^2 = 0.783$.

The mean Leverage (with SD) is: 0.057 (0.035)

We performed cross-validation on the data. Three leave-many-out procedures were applied, leaving out: a) 10%; b) 25%; and c) 50% of the training set. Each procedure was applied ten times.

The average q^2 (with Standard Deviation) were respectively:

10%: $q^2 = 0.707$ (0.230)

25%: $q^2 = 0.717$ (0.115)

50%: $q^2 = 0.772$ (0.037)

The QSAR model has not been assessed by the authors for its predictivity of the activity of external compounds. The external predictivity has been assessed by us on a set of amines retrieved from the open literature, and not used by the authors, with the following result:

External set, n = 33;

Correlation between experimental and predicted potency ($\log TA_{98}$) = 0.41;

Percentage of chemicals correctly predicted within 1 $\log TA_{98}$ unit: 0.36.

Characterization of the test set, in relation to the applicability domain of the training set

Structural domain: The basic structures of the test set chemicals were checked, and are within the range of the training set.

Range of descriptors values of the test set:

$\log P$: -0.61 6.71;

HOMO: -9.312 -7.861;

LUMO: -1.484 0.707;

Range of chemical similarity indices (PCs):

	Training set (N=88)		Test set (N=33)	
	Minimum	Maximum	Minimum	Maximum
Factor1	-1.292	1.424	-1.145	3.199
Factor2	-1.469	2.199	-1.314	1.836
Factor3	-1.930	2.174	-2.041	1.293
Factor4	-2.330	1.860	-1.567	1.988

A mechanistic interpretation, if possible

Overall, the principal factor affecting the relative mutagenicity of the aminoarenes is their

hydrophobicity (logP). Mutagenicity increased with increasing HOMO values; this positive correlation is in agreement with the known mechanism of action, because compounds with higher HOMO values are easier to oxidize and should be readily bioactivated. For the negative correlation with LUMO, no simple explanation could be offered by the authors, but the same evidence was observed by other authors on different sub-sets of chemicals and relative to other end-points.

2.4.1.2 QSAR 2, Ref. 4 in Appendix 1, Debnath et al., 1992, mutagenic potency in *Salmonella typhimurium* TA100

A defined endpoint

This QSAR is associated with a defined toxicity endpoint (genetic mutation), addressed by an officially recognised test method (Method B.13/14 Mutagenicity – Reverse Mutation test using bacteria – Annex V to *Directive 67/548/EEC*).

An unambiguous algorithm

The data and QSAR models are reported in Ref. 4. The mutagenic potency in TA100 strain (+ S9 activation system) was modelled by:

$$\log \text{TA100} = 0.92(\pm 0.23) \log P + 1.17(\pm 0.83)\text{HOMO} - 1.18(\pm 0.44)\text{LUMO} + 7.35(\pm 6.9)$$

$$n = 67, r = 0.877 (r^2 = 0.769), s = 0.708, F_{1,65} = 99.23$$

The mutagenic potency (log TA100) is expressed as log (revertants/nmol). High TA100 values indicate high mutagenic potency. The AM1 molecular orbital energies (HOMO and LUMO) are given in eV.

A defined domain of applicability

The applicability domain of the model is defined explicitly by the authors in terms of the structures to which it applies. The chemical set spans a large range of basic structures (aniline, biphenyl, anthracene, phenanthrene, fluorene, pyrene, fluoranthene, chrysene, quinoline, carbazole, phenazine) with one or two amine groups attached.

A number of outliers were excluded from the analysis, but they do not point to general rules for defining the applicability domain.

The ranges of the chemical descriptors are not reported explicitly, and were derived by us:

logP:	1.16	4.98
LUMO:	-1.330	0.702
HOMO:	-8.695	-7.528

Appropriate measures of goodness-of-fit, robustness and predictivity

The goodness-of-fit reported by the authors is the correlation coefficient. The data have been re-analysed by us. A Multiple Linear Regression (MLR) analysis reproduced the original QSAR equation, with

$$r^2 = 0.771;$$

$$\text{Adjusted } r^2 = 0.761;$$

$$q^2 = 0.740.$$

The mean Leverage (with SD) was: 0.060 (0.039)

We performed cross-validation on the data. Three leave-many-out procedures were applied, leaving out: a) 10%; b) 25%; and c) 50% of the training set. Each procedure was applied ten times.

The average q^2 (with Standard Deviation) were respectively:

$$10\%: q^2 = 0.657 (0.143)$$

$$25\%: q^2 = 0.740 (0.053)$$

$$50\%: q^2 = 0.705 (0.007)$$

The QSAR model has not been assessed by the authors for its predictivity of the activity of external compounds. Thus the external predictivity has been assessed by us on a set of amines retrieved from the open literature, and not used by the authors, with the following result:

External set, n = 29;

Correlation between experimental and predicted potency ($\log TA_{100}$) = 0.68;

Percentage of chemicals correctly predicted within 1 $\log TA_{100}$ unit = 0.57.

Characterization of the test set, in relation to the applicability domain of the training set

Structural domain: The basic structures of the test set chemicals were checked, and are within the range of the training set.

Range of descriptors values of the test set:

logP: 0.09 5.12
LUMO: -0.850 0.684
HOMO: -8.945 -7.861

Range of chemical similarity indices (PCs):

	Training set (N=67)		Test set (N=29)	
	Minimum	Maximum	Minimum	Maximum
Factor1	-1.097	0.952	-0.783	3.375
Factor2	-1.319	2.213	-1.157	1.092
Factor3	-1.415	2.347	-0.573	2.310
Factor4	-1.883	2.504	-1.773	1.958

A mechanistic interpretation, if possible

The model is very similar to that derived for the TA98 mutagenicity.

TA100 QSAR lacks the IL term present in the TA98 model. The authors hypothesize that larger amines are more capable of inducing frameshift mutations (TA98 is specific for frame-shift mutations, whereas TA100 is specific for base-pairs substitution mutations), and that this effect is not accounted for by the increase of $\log P$ for increasing sizes of the molecules.

2.4.1.3 QSAR 3, Ref. 23 in Appendix 1, Benigni et al., 2000, carcinogenic potency in mouse

A defined endpoint

This QSAR is associated with a defined toxicity endpoint (carcinogenicity), addressed by an officially recognised test method (Method B.32 Carcinogenicity test – Annex V to Directive 67/548/EEC).

An unambiguous algorithm

The data and QSAR models are reported in Ref. 23. The carcinogenic potency in mouse was modelled by:

$$\begin{aligned} \text{BRM} = & 0.88(\pm 0.27) \log P \times I(\text{monoNH}_2) + 0.29(\pm 0.20) \log P \times I(\text{diNH}_2) + 1.38(\pm 0.76) \text{HOMO} - \\ & 1.28(\pm 0.54) \text{LUMO} - 1.06(\pm 0.34) \Sigma \text{MR}_{2,6} - 1.10(\pm 0.80) \text{MR}_3 - 0.20(\pm 0.16) E_s(\text{R}) \\ & + 0.75(\pm 0.75) I(\text{diNH}_2) + 11.16(\pm 6.68) \\ n = & 37, r = 0.907, r^2 = 0.823, s = 0.381, F = 16.3, P < 0.001 \end{aligned}$$

where $\text{BRM} = \log(\text{MW}/\text{TD}_{50})_{\text{mouse}}$. TD_{50} is the daily dose required to halve the probability for an experimental animal of remaining tumorless to the end of its standard life span. High BRM values indicate high carcinogenic potency.

$\Sigma \text{MR}_{2,6}$ = sum of molar refractivity of substituents in the *ortho*-positions of the aniline ring;

MR_3 = molar refractivity of substituents in the *meta*-position of the aniline ring;

$E_s(\text{R})$ = Charton's substituent constant for substituents at the functional amino group;

$I(\text{monoNH}_2) = 1$ for compounds with only one amino group;

$I(\text{diNH}_2) = 1$ for compounds with more than one amino group.

MR and $E_s(\text{R})$ are tabulated values, from (Hansch, Leo & Hoekman 1995).

The PM3 (erroneously reported as AM1 in the paper) molecular orbital energies for HOMO and LUMO are given in eV.

A defined domain of applicability

The applicability domain of the model is defined explicitly by the authors in terms of the structures to which it applies. The chemical set spans a range of basic structures (anilines, biphenylamines, naphthylamines, aminofluorenes) with one or two amino groups attached.

The ranges of the chemical descriptors are not reported explicitly, and were derived by us:

logP	0.20	4.25
HOMO	-10.265	-7.990
LUMO	-1.997	0.438
$\Sigma MR_{2,6}$	0.180	2.640
MR ₃	0.100	0.800
$E_S (R)$	0	6.00

Appropriate measures of goodness-of-fit, robustness and predictivity

The goodness-of-fit reported by the authors is the correlation coefficient. The data have been re-analysed by us. A Multiple Linear Regression (MLR) analysis reproduced the original QSAR equation, with

$$r^2 = 0.821;$$

$$\text{Adjusted } r^2 = 0.767;$$

$$q^2 = 0.585.$$

The mean Leverage (with STD) was: 0.250 (0.157)

We performed cross-validation on the data. Three leave-many-out procedures were applied,

leaving out: a) 10%; b) 25%; and c) 50% of the training set. Each procedure was applied ten times.

The average q^2 (with Standard Deviation) were respectively:

10%: q^2 = negative

25%: q^2 = 0.213 (0.881)

50%: q^2 = 0.268 (0.321)

The QSAR model has not been assessed by the authors for its predictivity of the activity of external compounds. Thus the external predictivity has been assessed by us on a set of amines retrieved in the ISSCAN and CPDB databases, and not used by the authors, with the following result:

External set, $n = 12$;

Correlation between experimental and predicted potency (BRM) = 0.56;

Percentage of chemicals correctly predicted within 1 log unit (BRM) = 0.58.

Characterization of the test set, in relation to the applicability domain of the training set

Structural domain: The basic structures of the test set chemicals were checked, and are within the range of the training set.

Range of descriptors values of the test set:

logP	1.34	4.57
HOMO	-9.220	-8.043
LUMO	-1.043	0.411
$\Sigma MR_{2,6}$	0.200	0.660
MR_3	0.100	0.560
$E_S(R)$	0	4.000

Range of chemical similarity indices (PCs):

	Training set (N=36)		Test set (N=12)	
	Minimum	Maximum	Minimum	Maximum
Factor1	-1.157	1.890	-0.918	1.816
Factor2	-1.818	2.119	-1.653	2.176
Factor3	-1.068	2.622	-0.908	1.586

A mechanistic interpretation, if possible

The model is in agreement with the knowledge on the mechanism of action of the aromatic amines, whose major metabolic pathway requires oxidative activation of the chemicals. Moreover, the model agrees with those for the mutagenic potency in TA98 and TA100 Salmonella strains, with the same parameters in the same order of importance: hydrophobicity, HOMO, LUMO, and then a number of steric factors more specific for each individual experimental system.

2.4.1.4 QSAR 4, Ref. 23 in Appendix 1, Benigni et al., 2000, carcinogenic potency in rat

A defined endpoint

This QSAR is associated with a defined toxicity endpoint (carcinogenicity), addressed by an officially recognised test method (Method B.32 Carcinogenicity test – Annex V to Directive 67/548/EEC).

An unambiguous algorithm

The data and QSAR models are reported in Ref. 23. The carcinogenic potency in rat was modelled by:

$$\begin{aligned} \text{BRR} &= 0.35(\pm 0.18) \log P + 1.93(\pm 0.48) I(\text{Bi}) + 1.15(\pm 0.60) I(\text{F}) - 1.06(\pm 0.53) I(\text{BiBr}) \\ &+ 2.75(\pm 0.64) I(\text{RNNO}) - 0.48(\pm 0.30) \\ n &= 41, r = 0.933, r^2 = 0.871, s = 0.398, F = 47.4, P < 0.001 \end{aligned}$$

where $\text{BRR} = \log(\text{MW}/\text{TD50})_{\text{rat}}$. TD50 is the daily dose required to halve the probability for an experimental animal of remaining tumorless to the end of its standard life span. High BRR values indicate high carcinogenic potency.

$I(\text{Bi}) = 1$ for biphenyls;

$I(\text{BiBr}) = 1$ for biphenyls with a bridge between the phenyl rings;

$I(\text{RNNO}) = 1$ for compounds with the group N(Me)NO (nitroso group, with a methyl substitution at the amino nitrogen);

$I(\text{F}) = 1$ for fluoroamines.

A defined domain of applicability

The applicability domain of the model is defined explicitly by the authors in terms of the structures to which it applies. The chemical set spans a range of basic structures (anilines, biphenylamines, naphthylamines, aminofluorenes) with one or two amino groups attached.

The ranges of the chemical descriptors are not reported explicitly, and were derived by us:

logP: 0.23 3.73

Appropriate measures of goodness-of-fit, robustness and predictivity

The goodness-of-fit reported by the authors is the correlation coefficient. The data have been re-analysed by us. A Multiple Linear Regression (MLR) analysis reproduced the original QSAR equation, with

$r^2 = 0.871$;

Adjusted $r^2 = 0.852$;

$q^2 = 0.806$.

The mean Leverage (with STD) was: 0.146 (0.118)

We performed cross-validation on the data. Three leave-many-out procedures were applied, leaving out: a) 10%; b) 25%; and c) 50% of the training set. Each procedure was applied ten times. The average q^2 (with Standard Deviation) were respectively:

10%: $q^2 = 0.785$ (0.261)

25%: $q^2 = 0.721$ (0.325)

50%: $q^2 = 0.536$ (0.255)

The QSAR model has not been assessed by the authors for its predictivity of the activity of external compounds. Thus the external predictivity has been assessed by us on a set of amines retrieved in the ISSCAN and CPDB databases, and not used by the authors, with the following result:

External set, $n = 7$;

Correlation between experimental and predicted potency (BRR) = 0.48;

Percentage of chemicals correctly predicted within 1 log unit (BRR) = 0.71.

Characterization of the test set, in relation to the applicability domain of the training set

Structural domain: The basic structures of the test set chemicals were checked, and are within the range of the training set.

Range of descriptors values of the test set:

logP : 1.83 4.57

Range of chemical similarity indices (PCs):

	Training set (N=41)		Test set (N=7)	
	Minimum	Maximum	Minimum	Maximum
Factor1	-1.155	1.887	-1.015	1.643
Factor2	-1.754	1.797	-1.271	1.353
Factor3	-2.373	1.774	-0.651	0.973

A mechanistic interpretation, if possible

The logP and steric factors in the model can be interpreted in mechanistic terms, and appear also in the models for the carcinogenic potency in mouse and for the mutagenic potency. However, the electronic reactivity terms HOMO and LUMO present in the above equations are missing in this equation.

2.4.1.5 QSAR 5, our unpublished results, mutagenic activity in *Salmonella typhimurium* TA98

A defined endpoint

This QSAR is associated with a defined toxicity endpoint (genetic mutation), addressed by an officially recognised test method (Method B.13/14 Mutagenicity – Reverse Mutation test using bacteria – Annex V to *Directive 67/548/EEC*).

An unambiguous algorithm

The data on the mutagenic activity on the TA98 strain are in Ref. 4. The QSAR model was generated expressly by us for this project. The mutagenicity data (yes/no) were analysed with Canonical Discriminant Analysis. The mutagenic activity was modelled by:

$$w = -0.32 \text{ HOMO} + 0.97 \text{ LUMO} - 0.28 \text{ MR}_5 + 0.27 \text{ MR}_3 + 0.50 \text{ MR}_6$$

$$w(\text{mean,Class1}) = -0.31 \quad N1 = 21$$

$$w(\text{mean,Class2}) = 1.39 \quad N2 = 94$$

where N1 = number of non-mutagens (Class 1) and N2 = number of mutagens (Class 2). The threshold is the midpoint between w(mean,Class1) and w(mean,Class2).

The AM1 molecular orbital energies for HOMO and LUMO are given in eV. MR₅, MR₃ and MR₆ are the MR contributions of substituents in position 3, 5, and 6 to the amino group.

The coefficients in the equation are standardized, so they reflect the relative importance of the descriptors in the discrimination. To apply this equation to an external test set, one has to standardize the values of the descriptors according to the training set (i.e., subtract the mean and divide by the SD of the training set).

The equation correctly reclassified 82.6% (Accuracy) of the compounds (Class1, nonmutagens , 95.2% (Specificity; Class2, mutagens, 79.8% (Sensitivity)).

Squared Canonical Correlation = 0.31

A defined domain of applicability

The applicability domain of the model can be inferred from the structures to which it applies. The chemical set spans a large range of basic structures (aniline, biphenyl, anthracene, phenanthrene, fluorene, pyrene, fluoranthene, chrysene, quinoline, carbazole, phenazine) with one or two amino groups attached.

The ranges of the chemical descriptors are:

HOMO:	-10.020	-7.528
LUMO:	-1.691	0.722
MR ₅	0.000	0.800
MR ₃	0.100	3.170
MR ₆	0.056	1.500

Appropriate measures of goodness-of-fit, robustness and predictivity

The goodness-of-fit of the model is reported above (accuracy, sensitivity and specificity).

We also performed cross-validation on the data. Three leave-many-out procedures were applied, leaving out: a) 10%; b) 25%; and c) 50% of the training set. In addition, each procedure was applied in two different ways, by generating test sets with the following characteristics: 1) Option 1: with

the same proportion Class1/Class2 present in the whole sample of chemicals; 2) Option 2: without the above constraint. Each procedure was applied ten times.

Option 1

	Sensitivity	Specificity	Accuracy
-10%:	81.1	90.0	82.7
-25%:	77.1	96.0	80.3
-50%:	77.9	84.5	79.1

Option 2

	Sensitivity	Specificity	Accuracy
-10%:	82.6	100.0	86.7
-25%:	79.1	91.5	81.4
-50%:	75.9	83.0	76.7

The QSAR model has been assessed by us for its predictivity of the activity of external compounds (retrieved in the open literature), with the following result:

External set, n = 54;

Percentage of chemicals correctly predicted = 0.63

Characterization of the test set, in relation to the applicability domain of the training set

Structural domain: The basic structures of the test set chemicals were checked, and are within the range of the training set.

Range of descriptors values of the test set:

HOMO:	-10.212	-7.861
LUMO:	-2.028	0.709
MR ₅	0.100	0.740
MR ₃	0.100	0.800
MR ₆	0.100	1.500

Range of chemical similarity indices (PCs):

	Training set (N=115)		Test set (N=54)	
	Minimum	Maximum	Minimum	Maximum
Factor1	-1.317	0.997	-1.203	3.085
Factor2	-1.452	2.070	-1.414	1.724
Factor3	-1.721	2.490	-1.543	0.957
Factor4	-2.085	2.461	-2.023	1.609

A mechanistic interpretation, if possible

Whereas the principal factor that affects the relative mutagenicity (potency) of the aminoarenes is their hydrophobicity (logP), followed by electronic factors (HOMO and LUMO) and then steric factors, the model for the yes/no activity shows no influence of logP. This indicates that the potential to be active depends on a threshold of reactivity (HOMO and LUMO), and on the steric hindrance at substitution positions 3, 5, and 6 of the ring. The parameters in the model are mechanistically linked to requirements for metabolic activation.

2.4.1.6 QSAR 6, our unpublished results, mutagenic activity in *Salmonella typhimurium* TA100

A defined endpoint

This QSAR is associated with a defined toxicity endpoint (genetic mutation), addressed by an officially recognised test method (Method B.13/14 Mutagenicity – Reverse Mutation test using bacteria – Annex V to *Directive 67/548/EEC*).

An unambiguous algorithm

The data on the mutagenic activity on the TA100 strain are in Ref. 4. The QSAR model was generated expressly by us for this project. The mutagenicity data (yes/no) were analysed with Canonical Discriminant Analysis. The mutagenic activity was modelled by:

$$w = -0.65 \text{ HOMO} + 0.69 \text{ LUMO} + 0.39 \text{ MR}_2 + 0.39 \text{ MR}_3 + 0.46 \text{ MR}_6$$

$$w(\text{mean,Class1}) = 1.04 \text{ N1} = 43$$

$$w(\text{mean,Class2}) = -0.61 \text{ N2} = 73$$

where N1 = number of non-mutagens (Class 1) and N2 = number of mutagens (Class 2). The threshold is the midpoint between w(mean,Class1) and w(mean,Class2).

The AM1 molecular orbital energies for HOMO and LUMO are given in eV. MR₂, MR₃, MR₆ are the MR contributions of substituents in position 2, 3, and 6 to the amino group.

The coefficients in the equation are standardized, so they reflect the relative importance of the descriptors in the separation. To apply this equation to an external test set, one has to standardize the values of the descriptors according to the training set (i.e., subtract the mean and divide by the SD of the training set).

The equation correctly reclassified 78.6% (Accuracy) of the compounds (Class1, nonmutagens, 88.6% (Specificity); Class2, mutagens, 72.6% (Sensitivity)).

Squared Canonical Correlation: 0.39.

A defined domain of applicability

The applicability domain of the model can be inferred from the structures to which it applies. The chemical set spans a large range of basic structures (aniline, biphenyl, anthracene, phenanthrene, fluorene, pyrene, fluoranthene, chrysene, quinoline, carbazole, phenazine) with one or two amino groups attached.

The ranges of the chemical descriptors are:

HOMO	-9.032	-7.528
LUMO	-1.330	0.722
MR ₂	0.090	2.980
MR ₃	0.100	2.980
MR ₆	0.056	1.500

Appropriate measures of goodness-of-fit, robustness and predictivity

The goodness-of-fit of the model is reported above (accuracy, sensitivity and specificity). We also performed cross-validation on the data. Three leave-many-out procedures were applied, leaving out: a) 10%; b) 25%; and c) 50% of the training set. In addition, each procedure was applied in two different ways, by generating test sets with the following characteristics: 1) Option 1: with the same proportion Class1/Class2 present in the whole sample of chemicals; 2) Option 2: without the above constraint. Each procedure was applied ten times.

Option 1:

	Sensitivity	Specificity	Accuracy
-10%:	57.1	80.0	65.5
-25%:	71.7	93.0	79.3
-50%:	73.2	80.9	76.1

Option 2:

	Sensitivity	Specificity	Accuracy
-10%:	69.7	96.8	78.3
-25%:	68.4	86.6	75.9
-50%:	71.2	80.5	74.7

The QSAR model has been assessed by us for its predictivity of the activity of external compounds (retrieved in the open literature), with the following result:

External set, n = 52;

Percentage of chemicals correctly predicted = 0.69

Characterization of the test set, in relation to the applicability domain of the training set

Structural domain: The basic structures of the test set chemicals were checked, and are within the range of the training set.

Range of descriptors values of the test set:

HOMO	-10.212	-7.861
LUMO	-2.028	0.709
MR ₂	0	1.960
MR ₃	0.100	0.800
MR ₆	0.100	1.500

Range of chemical similarity indices (PCs):

	Training set (N=116)		Test set (N=52)	
	Minimum	Maximum	Minimum	Maximum
Factor1	-1.078	1.011	-0.969	3.109
Factor2	-1.600	1.681	-1.536	1.537
Factor3	-2.064	1.865	-2.125	0.730
Factor4	-1.296	2.522	-1.499	1.302
Factor5	-1.722	2.049	-1.063	2.251

A mechanistic interpretation, if possible

Whereas the principal factor that affects the relative mutagenicity (potency) of the aminoarenes is their hydrophobicity (logP), followed by electronic factors (HOMO and LUMO) and then steric factors, the model for the yes/no activity shows no influence of logP. This indicates that the potential to be active depends on a threshold of reactivity (HOMO and LUMO), and on the steric hindrance at substitution positions 2, 3, and 6 of the ring. The parameters in the model are mechanistically linked to requirements for metabolic activation.

2.4.1.7 QSAR 7, Ref. 24 in Appendix 1 (Eq. 4 in the paper), Franke et al., 2001, carcinogenic activity in rodents (overall)

A defined endpoint

This QSAR is associated with a defined toxicity endpoint (carcinogenicity), addressed by an officially recognised test method (Method B.32 Carcinogenicity test – Annex V to Directive 67/548/EEC).

An unambiguous algorithm

The data and the QSAR model is in Ref. 24, Eq. 4. The rodent carcinogenicity data (overall yes/no score from four experimental groups: rat, mouse, male, female) were analysed with Canonical Discriminant Analysis. The carcinogenicity was modelled by:

$$w = -2.86 L(R) + 2.65 B5(R) - 1.16 \text{HOMO} + 1.76 \text{LUMO} + 0.40 \text{MR}_3 + 0.58 \text{MR}_5 + 0.54 \text{MR}_6 - 1.55 \text{I(An)} + 0.74 \text{I(NO}_2) - 0.55 \text{I(BiBr)}$$

$$w(\text{mean,Class1}) = -1.56 \quad N1 = 13$$

$$w(\text{mean,Class2}) = 0.38 \quad N2 = 53$$

where N1 = number of non-carcinogens (Class 1) and N2 = number of carcinogens (Class 2). The threshold is the midpoint between w(mean,Class1) and w(mean,Class2).

L(R) (length) and B5(R) (maximal width) are Sterimol parameters (tabulated in (Verloop 1987)). The PM3 (erroneously AM1 in the paper) molecular orbital energies for HOMO and LUMO are given in eV. MR₃, MR₅, MR₆ are the MR contributions of substituents in position 3, 5, and 6 to the amino group. I(An), I(NO₂), and I(BiBr) are indicator variables that take value = 1 for anilines, for the presence of a NO₂ group, and for biphenyls with a bridge between the phenyl rings, respectively.

The coefficients in the equation are standardized, so they reflect the relative importance of the descriptors in the separation. Thus to apply this equation to an external test set, one has to standardize the values of the descriptors according to the training set (i.e., subtract the mean and divide by the SD of the training set).

The equation correctly reclassified 87.9% (Accuracy) of the compounds (Class1, noncarcinogens , 84.6% (Specificity); Class2, carcinogens, 88.7% (Sensitivity)).

Squared Canonical Correlation (our calculation): 0.38.

A defined domain of applicability

Regarding the applicability domain of the model, the original paper lists the basic substructures to which it applies (aniline, biphenyl, naphthalene, fluorene) with one or two amino groups attached. The ranges of chemical descriptors values are not reported explicitly, and were determined by us:

L(R):	2.06	5.97
B5(R):	1.00	4.04
HOMO:	-9.544	-7.989
LUMO:	-1.594	0.438
MR ₃ :	0.10	0.80
MR ₅ :	0.09	1.49
MR ₆ :	0.09	0.60

Appropriate measures of goodness-of-fit, robustness and predictivity

The goodness-of-fit reported by the authors consists of the accuracy, sensitivity and specificity parameters (see above). Our re-analysis of the original data reproduced the reported QSAR equation and statistical parameters.

We performed cross-validation on the data. Three leave-many-out procedures were applied, leaving out: a) 10%; b) 25%; and c) 50% of the training set. In addition, each procedure was applied in two different ways, by generating test sets with the following characteristics: 1) Option 1: with the same proportion Class1/Class2 present in the whole sample of chemicals; 2) Option 2: without the above constraint. Each procedure was applied ten times.

Option 1:

	Sensitivity	Specificity	Accuracy
-10%:	76.0	70.0	75.0
-25%:	79.2	70.0	77.5
-50%:	76.7	70.0	75.3

Option 2:

	Sensitivity	Specificity	Accuracy
-10%:	89.3	70.0	85.7
-25%:	80.4	82.5	80.0
-50%:	80.7	66.5	78.5

The QSAR model has not been assessed by the authors for its predictivity of the activity of external compounds. Thus the external predictivity has been assessed by us on a set of amines retrieved in the ISSCAN and CPDB databases, and not used by the authors, with the following result:

External set, $n = 27$;

Percentage of chemicals correctly predicted = 0.67

Characterization of the test set, in relation to the applicability domain of the training set

Structural domain:

The basic structures of the test set chemicals were checked, and are within the range of the training set.

Range of descriptors values of the test set:

L(R):	2.060	10.790
B5(R):	1.000	7.990
HOMO:	-9.416	-8.043
LUMO	-1.304	0.461
MR ₃ :	0.100	0.800
MR ₅ :	0.100	0.740
MR ₆ :	0.100	0.600

Range of chemical similarity indices (PCs):

	Training set (N=66)		Test set (N=28)	
	Minimum	Maximum	Minimum	Maximum
Factor1	-1.160	2.041	-1.020	1.952
Factor2	-1.678	2.007	-1.688	2.176
Factor3	-1.914	1.618	-2.016	1.777
Factor4	-1.892	1.997	-1.932	1.172

A mechanistic interpretation, if possible

Whereas the principal factor that affects the relative carcinogenicity (potency) of the aminoarenes is their hydrophobicity (logP), followed by electronic factors (HOMO and LUMO) and then steric factors, the model for the yes/no activity shows no influence of logP. This indicates that the potential to be active depends on a threshold of reactivity (HOMO and LUMO), and on the steric hindrance at substitution positions 3, 5, and 6 of the ring, together with steric hindrance due to bulky substituents to the nitrogen. The parameters in the model are mechanistically linked to requirements for metabolic activation.

2.4.1.8 QSAR 8, Ref. 24 (Eq. 5 in the paper) in Appendix 1, Franke et al., 2001, carcinogenic activity in rodents (overall)

A defined endpoint

This QSAR is associated with a defined toxicity endpoint (carcinogenicity), addressed by an officially recognised test method (Method B.32 Carcinogenicity test – Annex V to Directive 67/548/EEC).

An unambiguous algorithm

The data and the QSAR model is in Ref. 24, Eq. 5. The rodent carcinogenicity data (overall yes/no score from four experimental groups: rat, mouse, male, female) were analysed with Canonical Discriminant Analysis. The carcinogenicity was modelled by:

$$w = -3.42 L(R) + 3.11 B5(R) - 1.57 \text{HOMO} + 2.19 \text{LUMO} + 0.66 \text{MR}_3 + 0.65 \text{MR}_5 + 0.54 \text{MR}_6 - 1.64 I(\text{An}) + 0.57 I(\text{NO}_2) - 0.63 I(\text{BiBr})$$

$$w(\text{mean,Class1}) = -2.04 \quad N1 = 12$$

$$w(\text{mean,Class2}) = 0.47 \quad N2 = 52$$

where N1 = number of non-carcinogens (Class 1) and N2 = number of carcinogens (Class 2). The threshold is the midpoint between $w(\text{mean,Class1})$ and $w(\text{mean,Class2})$.

$L(R)$ (length) and $B5(R)$ (maximal width) are Sterimol parameters (tabulated in (Verloop 1987)). The PM3 (erroneously AM1 in the paper) molecular orbital energies for HOMO and LUMO are given in eV. MR_3 , MR_5 , MR_6 are the MR contributions of substituents in position 3, 5, and 6 to the amino group. $I(\text{An})$, $I(\text{NO}_2)$, and $I(\text{BiBr})$ are indicator variables that take value = 1 for anilines, for

the presence of a NO₂ group, and for biphenyls with a bridge between the phenyl rings, respectively.

The coefficients in the equation are standardized, so they reflect the relative importance of the descriptors in the separation. Thus to apply this equation to an external test set, one has to standardize the values of the descriptors according to the training set (i.e., subtract the mean and divide by the SD of the training set).

The equation correctly reclassified 93.7% (Accuracy) of the compounds (Class1, noncarcinogens, 92.7% (Specificity); Class2, carcinogens, 94.2% (Sensitivity)).

Squared Canonical Correlation (our calculation): 0.50.

This equation was obtained after exclusion from the training set of two compounds misclassified by QSAR7 (Eq. 4 in Ref. 24). The exclusion of the two chemicals improved both the goodness of fit and the external predictivity (see below).

A defined domain of applicability

Regarding the applicability domain of the model, the original paper lists the basic substructures to which it applies (aniline, biphenyl, naphthalene, fluorene) with one or two amino groups attached. The ranges of chemical descriptors values are not reported explicitly, and were determined by us:

L(R)	2.060	5.970
B5(R)	1.000	4.040
HOMO	-9.544	-7.990
LUMO	-1.594	0.438

MR ₃	0.100	0.800
MR ₅	0.090	1.490
MR ₆	0.090	0.600

Appropriate measures of goodness-of-fit, robustness and predictivity

The goodness-of-fit reported by the authors consists of the accuracy, sensitivity and specificity parameters (see above). Our re-analysis of the original data reproduced the reported QSAR equation and statistical parameters.

We performed cross-validation on the data. Three leave-many-out procedures were applied, leaving out: a) 10%; b) 25%; and c) 50% of the training set. In addition, each procedure was applied in two different ways, by generating test sets with the following characteristics: 1) Option 1: with the same proportion Class1/Class2 present in the whole sample of chemicals; 2) Option 2: without the above constraint. Each procedure was applied ten times.

Option 1:

	Sensitivity	Specificity	Accuracy
10%:	80.0	70.0	78.3
25%:	84.6	80.0	83.8
50%:	83.8	81.7	83.4

Option 2:

	Sensitivity	Specificity	Accuracy
10%:	85.2	71.7	81.7
25%:	79.9	76.2	78.1
50%:	82.5	73.1	80.3

The QSAR model has not been assessed by the authors for its predictivity of the activity of external compounds. Thus the external predictivity has been assessed by us on a set of amines retrieved in the ISSCAN and CPDB databases, and not used by the authors, with the following result:

External set, n = 27;

Percentage of chemicals correctly predicted = 0.70

Characterization of the test set, in relation to the applicability domain of the training set

Structural domain: The basic structures of the test set chemicals were checked, and are within the range of the training set.

Range of descriptors values of the test set:

L(R)	2.060	10.790
B5(R)	1.000	7.990
HOMO	-9.416	-8.043
LUMO	-1.304	0.461
MR ₃	0.100	0.800
MR ₅	0.100	0.740
MR ₆	0.100	0.600

Range of chemical similarity indices (PCs):

	Training set (N=64)		Test set (N=28)	
	Minimum	Maximum	Minimum	Maximum
Factor1	-1.157	2.021	-1.002	1.933
Factor2	-1.765	1.854	-1.742	2.017
Factor3	-1.852	1.836	-1.938	2.040
Factor4	-1.883	2.038	-1.946	1.154

A mechanistic interpretation, if possible

Whereas the principal factor that affects the relative carcinogenicity (potency) of the aminoarenes is their hydrophobicity (logP), followed by electronic factors (HOMO and LUMO) and then steric factors, the model for the yes/no activity shows no influence of logP. This indicates that the potential to be active depends on a threshold of reactivity (HOMO and LUMO), and on the steric hindrance at substitution positions 3, 5, and 6 of the ring, together with steric hindrance due to bulky substituents to the nitrogen. The parameters in the model are mechanistically linked to requirements for metabolic activation.

2.4.2 Nitroarenes

Nitroarenes are highly represented among the EU HPV's (Table II). For this class, only models for the mutagenic potency in *Salmonella typhimurium* are available in the literature. Beside revising two short listed models (Refs. 32 and 33 in Appendix 1), we found in the literature a set of further chemicals suitable for assessing their external predictivity: the results are reported below.

2.4.2.1 QSAR 9, Ref. 32, mutagenic potency in *Salmonella typhimurium* TA98

A defined endpoint

This QSAR is associated with a defined toxicity endpoint (genetic mutation), addressed by an officially recognised test method (Method B.13/14 Mutagenicity – Reverse Mutation test using bacteria – Annex V to *Directive 67/548/EEC*).

An unambiguous algorithm

The data and QSAR models are reported in Ref. 32. The mutagenic potency in TA98 strain (without S9 activation system) was modelled by:

$$\log \text{TA98} = 0.65(\pm 0.16) \log P - 2.90(\pm 0.59) \log (\beta 10^{\log P} + 1) - 1.38(\pm 0.25) \text{LUMO} \\ + 1.88 (\pm 0.39) I_1 - 2.89 I_a (\pm 0.81) - 4.15(\pm 0.58)$$

$$n = 188, r = 0.900 (r^2 = 0.810), s = 0.886, \log P_0 = 4.93, \log \beta = 5.48, F_{1,181} = 48.6$$

The mutagenic potency (log TA98) is expressed as log (revertants/nmol); increasing TA98 values correspond to increasing mutagenic potency. The AM1 molecular orbital energies are given in eV.

I_1 is an indicator variable that assumes a value of 1 for compounds with three or more fused rings;

I_a takes the value of 1 for five examples of acenaphthylenes.

A defined domain of applicability

The applicability domain of the model is defined by the authors in terms of the structures to which it applies. The chemical set spans a very large range of basic structures (e.g., benzene to coronene and many different types of heterocycles) which are listed in the original paper.

The ranges of the chemical descriptors are not reported explicitly, and were derived by us:

LUMO: -3.770 -0.530

logP. -0.02 7.84

Appropriate measures of goodness-of-fit, robustness and predictivity

The goodness-of-fit reported by the authors is the correlation coefficient and related statistics (see above). The data were re-analysed by us for this work. A Multiple Linear Regression (MLR) analysis reproduced the original QSAR equation, with

$r^2 = 0.811$;

Adjusted $r^2 = 0.805$;

$q^2 = 0.890$.

The Cross-Validated r^2 (q^2) is $= 1 - (\text{sum of squares of the predictive residuals} / \text{sum of squares of the mean-centered response data})$.

The mean Leverage (with STD) was: 0.032 (0.040)

We performed cross-validation on the data. Three leave-many-out procedures were applied, leaving out: a) 10%; b) 25%; and c) 50% of the training set. Each procedure was applied ten times.

The average q^2 (with Standard Deviation) were respectively:

10%: $q^2 = 0.795$ (0.064)

25%: $q^2 = 0.796$ (0.041)

50%: $q^2 = 0.788$ (0.023)

The QSAR model has not been assessed by the authors for its predictivity of the activity of external compounds. Thus the external predictivity has been assessed by us on a set of nitroarenes retrieved in the literature, and not used by the authors, with the following result:

External set, $n = 30$;

Correlation between experimental and predicted potency ($\log TA_{98}$) = -0.23;

Percentage of chemicals correctly predicted within 1 $\log TA_{98}$ unit = 0.43.

Characterization of the test set, in relation to the applicability domain of the training set

Structural domain: The basic structures of the test set chemicals were checked, and are within the range of the training set.

Range of descriptors values of the test set:

LUMO: -1.815 -0.870

$\log P$ 0.28 7.59

Range of chemical similarity indices (PCs):

	Training set (N=188)		Test set (N=30)	
	Minimum	Maximum	Minimum	Maximum
Factor1	-1.026	2.630	-0.931	0.893
Factor2	-1.547	2.853	-0.335	2.219
Factor3	-1.902	2.800	-1.565	2.027
Factor4	-2.456	1.418	-2.485	1.267

A mechanistic interpretation, if possible

Overall, the principal factor affecting the relative mutagenicity of the nitroarenes was their hydrophobicity (logP). Mutagenicity increased with decreasing LUMO values; this negative correlation is in agreement with the known mechanism of action, because compounds with lower LUMO values are easier to reduce and should be readily bioactivated.

The similarity with QSAR 10 (see *infra*) adds credibility to this model (lateral validation).

2.4.2.2 QSAR 10, Ref. 33, mutagenic potency in *Salmonella typhimurium* TA100

A defined endpoint

This QSAR is associated with a defined toxicity endpoint (genetic mutation), addressed by an officially recognised test method (Method B.13/14 Mutagenicity – Reverse Mutation test using bacteria – Annex V to *Directive 67/548/EEC*).

An unambiguous algorithm

The data and QSAR models are reported in Ref. 33. The mutagenic potency in TA100 strain (without S9 activation system) was modelled by:

$$\log \text{TA100} = 1.20(\pm 0.15) \log P - 3.40(\pm 0.74) \log (\beta 10^{\log P} + 1) - 2.05(\pm 0.32) \text{LUMO} \\ - 3.50(\pm 0.82) I_a + 1.86(\pm 0.74) I_{\text{ind}} - 6.39(\pm 0.73)$$

$$n = 117, r = 0.886 (r^2 = 0.785), s = 0.835, \log P_0 = 5.44, \log \beta = -5.7, F_{1,110} = 24.7$$

The mutagenic potency ($\log \text{TA100}$) is expressed as \log (revertants/nmol); increasing TA100 values correspond to increasing mutagenic potency. The AM1 molecular orbital energies are given in eV. I_{ind} is an indicator variable that assumes a value of 1 for six examples of 1- and 2-methylindazoles; I_a takes the value of 1 for five examples of acenaphthylenes.

A defined domain of applicability

The applicability domain of the model is defined by the authors in terms of the structures to which it applies. The chemical set spans a very large range of basic structures (e.g., benzene to coronene and many different types of heterocycles) which are listed in the original paper.

The ranges of the chemical descriptors are not reported explicitly, and were derived by us:

LUMO: -3.406 -0.690
logP: -0.47 7.84

Appropriate measures of goodness-of-fit, robustness and predictivity

The goodness-of-fit reported by the authors is the correlation coefficient and related statistics (see above). The data were re-analysed by us for this work. A Multiple Linear Regression (MLR) analysis reproduced the original QSAR equation, with

$$r^2 = 0.785;$$

$$\text{Adjusted } r^2 = 0.775;$$

$$q^2 = 0.768.$$

The Cross-Validated r^2 (q^2) is $= 1 - (\text{sum of squares of the predictive residuals} / \text{sum of squares of the mean-centered response data})$.

The mean Leverage (with STD) was: 0.051 (0.070)

We performed cross-validation on the data. Three leave-many-out procedures were applied, leaving out: a) 10%; b) 25%; and c) 50% of the training set. Each procedure was applied ten times. The average q^2 (with Standard Deviation) were respectively:

$$10\%: q^2 = 0.726 (0.514)$$

$$25\%: q^2 = 0.756 (0.058)$$

$$50\%: q^2 = 0.757 (0.033)$$

The QSAR model has not been assessed by the authors for its predictivity of the activity of external compounds. Thus the external predictivity has been assessed by us on a set of nitroarenes retrieved in the literature, and not used by the authors, with the following result:

External set, $n = 25$;

Correlation between experimental and predicted potency ($\log TA_{100}$) = 0.36;

Percentage of chemicals correctly predicted within 1 $\log TA_{100}$ unit = 0.32.

Characterization of the test set, in relation to the applicability domain of the training set

Structural domain: The basic structures of the test set chemicals were checked, and are within the range of the training set.

Range of descriptors values of the test set:

LUMO:	-1.830	-0.704
logP	0.23	7.59

Range of chemical similarity indices (PCs):

	Training set (N=117)		Test set (N=25)	
	Minimum	Maximum	Minimum	Maximum
Factor1	-1.571	1.706	-0.192	1.718
Factor2	-1.457	2.315	-0.619	1.109
Factor3	-1.514	3.234	-1.589	0.610
Factor4	-1.955	2.967	-2.391	0.030

A mechanistic interpretation, if possible

Overall, the principal factor affecting the relative mutagenicity of the nitroarenes was their hydrophobicity ($\log P$). Mutagenicity increased with decreasing LUMO values; this negative correlation is in agreement with the known mechanism of action, because compounds with lower LUMO values are easier to reduce and should be readily bioactivated.

The similarity with QSAR 9 adds credibility to this model (lateral validation).

2.4.3 Polycyclic aromatic hydrocarbons (PAH)

The PAH –as pure chemicals- are poorly represented among the EU HPV's (Table II), however they are widely diffused in mixtures (e.g., in petrol products). Only models for the (skin) carcinogenicity of the PAHs were retrieved in the literature. No external data sets, with a coherent measure of biological activity, were identified for assessing predictivity.

2.4.3.1 QSAR 11, Ref. 47 in Appendix 1, Zhang et al., 1992, Skin carcinogenicity in rodents

A defined endpoint

This QSAR is associated with a defined toxicity endpoint (carcinogenicity), addressed by an officially recognised test method (Method B.32 Carcinogenicity test – Annex V to Directive 67/548/EEC).

An unambiguous algorithm

The data and QSAR models are reported in Ref. 47. The skin carcinogenic potency in mice was modelled by:

$$\log I_{ball} = 0.55(\pm 0.09) \log P - 1.17(\pm 0.14) \log (.10 \log P + 1) + 0.39(\pm 0.11) LK \\ + 0.47(\pm 0.26) HOMO + 1.93(\pm 2.4)$$

$$n = 161, r = 0.845 (r^2 = 0.714), s = 0.350, \log . = -6.81, F_{1,155} = 12.8$$

where:

$$I_{ball} \text{ index} = (\text{tumor incidence}) (100\%) / (\text{mean latent period in days})$$

with

tumor incidence = (number of animal with tumors) / (number of animals alive when the first tumor appears);

LK is an indicator variable assigned a value 1 for all chemicals where a substituent is attached to a L or K region (Structure 1).

The electronic parameter HOMO was calculated with the AM1 procedure.

Structure 1: L, K and Bay regions of PAHs



A defined domain of applicability

The model applies to a very large range of structures (both homo and heterocyclic); details on the structural variations are in the original paper.

The ranges of chemical descriptors values are not reported explicitly in the paper, and are as follows:

logP:	3.30	11.01
HOMO:	-9.13	-7.54

Appropriate measures of goodness-of-fit, robustness and predictivity

The goodness-of-fit reported by the authors is the correlation coefficient and related statistics (see above). The data were re-analysed by us for this work. A Multiple Linear Regression (MLR) analysis reproduced the original QSAR equation, with

$$r^2 = 0.713;$$

$$\text{Adjusted } r^2 = 0.705;$$

$$q^2 = 0.689.$$

The mean Leverage (with SD) was: 0.031 (0.030)

We performed cross-validation on the data. Three leave-many-out procedures were applied, leaving out: a) 10%; b) 25%; and c) 50% of the training set. Each procedure was applied ten times.

The average q^2 (with Standard Deviation) were respectively:

$$10\%: q^2 = 0.687 (0.113)$$

$$25\%: q^2 = 0.712 (0.075)$$

$$50\%: q^2 = 0.663 (0.061)$$

The QSAR model has not been assessed by the authors for its predictivity of the activity of external compounds; however, we were not able to find in the literature a set of external chemicals suitable for assessing the predictivity of the model.

A mechanistic interpretation, if possible

The QSAR model is in agreement with the theories regarding K-region (9,10 bond in phenanthrene and analogs) and L-region (region between the C14 and C17 positions of phenanthrene) activation as being responsible for carcinogenicity of these compounds. A positive

coefficient of LK means that a substitution in an L or K region inhibits metabolism at these points and then leads to increased potency of these congeners, other factors being equal.

2.4.4 α,β -Unsaturated aliphatic aldehydes

Aliphatic aldehydes are widely diffused among the EU HPVs. Here we analyze models for the mutagenic potency, and mutagenic activity of the α,β -unsaturated aliphatic aldehydes (Refs 60 and 61). The external predictivity was assessed by the authors of the models.

2.4.4.1 QSAR 12, Ref. 60 in Appendix 1, Benigni et al., 2003, mutagenic potency in *Salmonella typhimurium* TA100

A defined endpoint

This QSAR is associated with a defined toxicity endpoint (genetic mutation), addressed by an officially recognised test method (Method B.13/14 Mutagenicity – Reverse Mutation test using bacteria – Annex V to *Directive 67/548/EEC*).

An unambiguous algorithm

The data and QSAR model are reported in Ref. 60. The mutagenic potency in TA100 strain (without S9 activation system) was modelled by:

$$\log\text{TA100} = -4.58430 \text{ LUMO} - 3.66205 \text{ MR} + 72.46140 \text{ C}_{\text{carb}} + 2.55239 \log\text{P} \\ + 13.09442 \text{ C}_{\beta} - 12.61592$$

$$n = 17; r^2 = 0.84; \text{ cross-validated } r^2 (q^2) = 0.40;$$

The mutagenic potency (log TA100) is expressed as log (revertants/ μmol). The PM3 molecular orbital energies are given in eV. C_{carb} and C_{β} are the partial charges on the carbonilic and β carbon atoms.

A defined domain of applicability

The applicability domain of the model is restricted to α,β -unsaturated aldehydes with both aliphatic and aromatic substitutions.

The ranges of the chemical descriptors are not reported explicitly, and were derived by us:

LUMO	-1.159	-0.101
MR	1.655	5.171
C _{carb}	0.307	0.333
logP	-0.010	2.742
C _{β}	-0.149	0.036

Appropriate measures of goodness-of-fit, robustness and predictivity

The goodness-of-fit reported by the authors is the correlation coefficient and related statistics (see above). The data were re-analysed by us for this work. A Multiple Linear Regression (MLR) analysis reproduced the original QSAR equation, with

$$r^2 = 0.841;$$

$$\text{Adjusted } r^2 = 0.767;$$

$$q^2 = 0.387.$$

The mean Leverage (with STD) was: 0.253 (0.190)

We performed cross-validation on the data. Three leave-many-out procedures were applied, leaving out: a) 10%; b) 25%; and c) 50% of the training set. Each procedure was applied ten times.

The average q^2 (with Standard Deviation) were respectively:

10%: q^2 = negative

25%: q_2 = negative

50%: q_2 = negative

Regarding external predictivity, in a subsequent paper (Ref. 61) more chemicals of the same class were tested: only two out five resulted to be mutagenic, so a test of external predictivity is not representative given the small size of the sample. We were not able to retrieve in the literature further external chemicals with test data.

A mechanistic interpretation, if possible

The mutagenic potency in TA100 increases with decreasing values of LUMO and MR. Low LUMO values indicate a high propensity to accept electrons, hence high electrophilic reactivity. MR parametrizes the bulkiness of the molecules; the smaller the molecules, the higher their capacity to interact. Moreover, the mutagenic potency is favored by increasing partial charges on both the carbonyl and β carbons, and by increasing hydrophobicity (logP). Overall, the result of the QSAR analysis is concordant with the scientific evidence on mechanism of action of these chemicals, which react as direct electrophiles to form adducts with DNA and proteins (with the β and carbonyl carbons being the points of attack).

2.4.4.2 QSAR 13, Refs. 60 and 61 in Appendix 1, Benigni et al., 2003; 2005, mutagenic activity in *Salmonella typhimurium* TA100

A defined endpoint

This QSAR is associated with a defined toxicity endpoint (genetic mutation), addressed by an officially recognised test method (Method B.13/14 Mutagenicity – Reverse Mutation test using bacteria – Annex V to *Directive 67/548/EEC*).

An unambiguous algorithm

The data on the mutagenic activity on the TA100 strain are in Ref. 60. The mutagenicity data (yes/no) were analysed with Stepwise Linear Discriminant Analysis. The mutagenic activity was modelled by equations relative to negatives and positives, respectively:

$$w_{\text{negative}} = -47.13 + 38.25 \text{ MR} - 31.78 \log P + 30.47 \text{ LUMO}$$

$$w_{\text{positive}} = -20.52 + 25.42 \text{ MR} - 21.45 \log P + 19.78 \text{ LUMO}$$

$$n.\text{negatives} = 3; n.\text{positives} = 17.$$

The PM3 molecular orbital energies are given in eV.

To estimate the activity of external chemicals, the two equations are applied and the chemical is assigned to the class for which the resulting w value is highest.

The equation correctly reclassified 100% of the compounds; a Leave-One-Out cross-validation reclassified correctly 85% of the compounds.

Squared Canonical Correlation = 0.61.

A defined domain of applicability

The applicability domain of the model is restricted to α,β -unsaturated aldehydes with both aliphatic and aromatic substitutions.

The ranges of chemical descriptors values are not reported explicitly in the paper, and are as follows:

MR:	1.655	5.171
logP	-0.01	2.95
LUMO	-1.159	-0.101

Appropriate measures of goodness-of-fit, robustness and predictivity

The goodness-of-fit reported by the authors consists of the accuracy, sensitivity and specificity parameters (see above). Our re-analysis of the original data reproduced the reported QSAR equation and statistical parameters (including the LOO analysis).

Since the training set includes only 3 negatives, no meaningful LMO cross-validation is possible.

The QSAR model was assessed by the authors for its predictivity in Ref. 61. Five untested chemicals were identified in commercial catalogues, and then tested experimentally in the same laboratory. The predictivity of the model on the external test set was:

External set, $n = 5$;

Percentage of chemicals correctly predicted = 1.00.

Characterization of the test set, in relation to the applicability domain of the training set

Structural domain: The basic structures of the test set chemicals were checked, and are within the range of the training set.

Range of descriptors values of the test set:

MR	2.580	6.910
logP	0.59	3.49
LUMO	-0.827	0.074

Range of chemical similarity indices (PCs):

	Training set (N=20)		Test set (N=5)	
	Minimum	Maximum	Minimum	Maximum
Factor1	-0.976	1.524	-0.866	1.753
Factor2	-1.707	1.853	0.005	2.764
Factor3	-2.616	1.238	-2.508	1.717

A mechanistic interpretation, if possible

The mutagenic activity in TA100 depends on the same factors (LUMO, MR, logP) that influence the mutagenic potency of active compounds (high electrophilicity, low steric hindrance and high lipophilicity favor the activity). This result is concordant with the scientific evidence on mechanism of action of these chemicals, which react as direct electrophiles to form adducts with DNA and proteins (the β - and carbonyl carbons are the points of attack). The partial charges on the carbonyl and β carbons, that are present in the equation for potency (QSAR 12), do not appear to influence the mutagenic activity.

3 Non-local (Q)SARs

To respond to the lack of individual QSARs for many chemical classes, a series of non-local models for noncongeneric sets of chemicals, i.e., general prediction models hopefully able to cope with the thousands of chemicals present in the environment, have been generated (Benigni 2005; Benigni & Richard 1998). Among those, a special place is held by the models based on Structural Alerts (SA). These were originally created as compilations of the scientific knowledge on the mechanisms of chemical carcinogenicity, without any use of statistics; more recently, refinements have been attempted with the support of more formal approaches (e.g., statistics / artificial intelligence). The knowledge on the action mechanisms as exemplified by the SAs is routinely used in SAR assessment in the regulatory context. In addition, the SAs are at the basis of popular commercial systems (e.g., DEREK).

3.1 The short listed models

In the context of this evaluation, it should be noted that the SA-based models in Appendix 1 are potentially suited for in depth analyses, whereas other non-local models (listed in Appendix 1 as well) require the use of proprietary descriptors / algorithms (see also Appendix 2). Four SA models were identified as particularly promising, and were characterized in this study (Table I). They are very different in nature, ranging from completely non-formalized ones (expert knowledge, no use of statistics), to models formalized and implemented into computer software.

A first model is the compilation of SAs by John Ashby (Ref. 62) (see also (Ashby & Tennant 1988)). The latter reference includes additional SAs in respect to the classical poly-carcinogen reproduced in Figure 4a (e.g., PAH), as well as some detoxifying chemical functionalities (e.g., sulfonic groups on azo-dyes, sterically hindering groups on the aromatic amino nitrogen). This model has a total of 19 SAs.

The compilation of SAs by Bailey et al, (Ref. 65) was generated for being used in the regulatory context of the newly implemented Food and Contact Notification program of the U.S. Food and Drug Administration (FDA) Office for Food Additive Safety. The list of SAs is based on the Ashby's SAs, and on a related list compiled by Munro (Munro, Ford, Kennepohl & Sprenger 1996). It consists of 33 SAs.

Kazius et al., 2005, (Ref. 63) produced another list of SAs (29 in total), based on a computerized data mining analysis whose results were "supervised" with an eye to the expert knowledge formalized by John Ashby. As noted above, the Ashby's SAs are tailored on the mechanistic knowledge on chemical carcinogens, mainly restricted to the genotoxic (DNA reactive) carcinogens. The exercise by Kazius et al. 2005 used a mutagenicity database (4337 mutagens and nonmutagens from the Toxnet database <http://toxnet.nlm.nih.gov/>). Thus, the resulting SAs are typical of *Salmonella* mutagens, and for this reason they are rigorously restricted to the genotoxic carcinogens.

The fourth set of SAs was generated by Kazius et al 2006, (Ref. 64) in an exercise aimed at experimenting a new way of representing the chemicals (hierarchical graphs) and a new searching algorithm (called Gaston). The goal was to generate automatically SAs through artificial intelligence methods solely. This effort resulted in 6 "complex" SAs.

3.2 Building the capacity to manipulate the SAs

In order to be able to properly assess the short listed models, it is necessary to have the capacity to manipulate the SAs: thus, for the present work we built in our laboratory the capacity to implement the SAs into computer programs. In practice, we coded the SAs into .mol files, and we managed them with the Leadscape software (Leadscape Inc, Columbus, OH; <http://www.leadscope.com>) that is able to read Chemical Relational Databases and to perform substructure searching. Leadscape was used to manage the databases employed as probes as well.

However simple in principle, the coding of the SAs presents a range of difficulties. First of all, the SAs are usually defined in a quite generic way by the authors, without the degree of precision required by the computer programs. Thus, it is necessary an effort to interpret the definitions, and then assess the influence of various, slightly different ways of coding the SAs by checking how they perform on selected databases of known toxic / nontoxic chemicals. In practice we wrote standard .mol files of the SAs, but subsequently we had to modify some of them manually. Moreover, some SAs actually consist of a range of co-existing substructures; such SAs are difficult to implement into one file, and (when necessary) were replaced by sets of queries. An additional difficulty derives from the fact that different software programs have different idiosyncrasies, and the .mol files written for one program may be not appropriate for another program. The result is that an accurate coding of the SAs, taking into account both the intentions of the authors and the idiosyncrasies of the software programs, is quite a delicate and time consuming task. In spite of the above difficulties, we were able to code the four sets of SAs mentioned above, and to check their performance.

3.3 Probes for the SAs: the databases

The models selected were compared by checking their ability to identify mutagens and carcinogens in different databases. Three databases with general relevance were used as probes. It should be remarked that the three databases contain chemicals with very diverse structures; a majority of them has been tested for mutagenicity and carcinogenicity because of their environmental importance, however they also contain many pharmaceuticals tested for their effects on human health.

One is the database of *Salmonella* mutagenicity data collected by (Kazius, McGuire & Bursi 2005) for their refinement of the Ashby' SAs. It consists of 4337 mutagens and nonmutagens, that the authors retrieved mainly from the Toxnet public database (<http://toxnet.nlm.nih.gov/>).

The other two databases include both rodent carcinogenicity and *Salmonella* mutagenicity data. One is the Carcinogenic Potency DataBase (CPDB) hosted at the DSSTox website (<http://www.epa.gov/ncct/dsstox/index.html>) (n = 1189), and the other one is the ISSCAN database on animal carcinogens, hosted at the website of the Istituto Superiore di Sanita' (<http://www.iss.it/ampp/dati/cont.php?id=233&lang=1&tipo=7>) (n = 890). The latter two databases are largely overlapping; however, there are some differences both in the chemicals included and in a number of mutagenicity and carcinogenicity calls (Benigni & Bossa 2006; Richard 2004) (Richard 2004) (Richard & Williams 2003).

3.4 The characterization of the SAs-based models

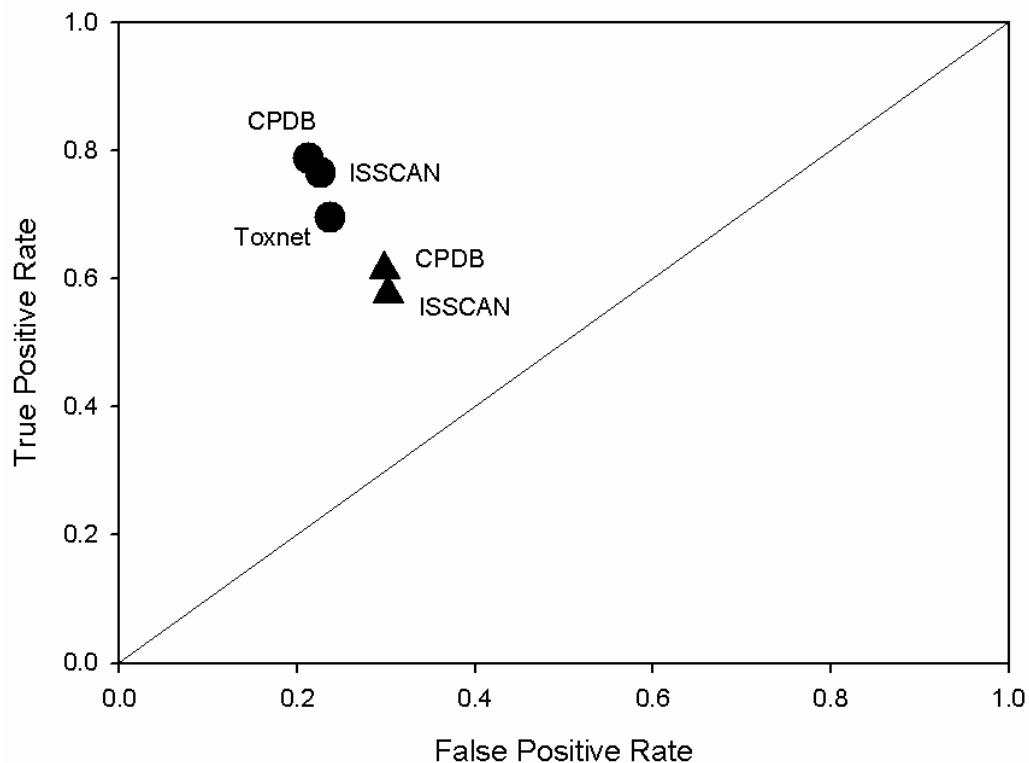
Since two out four models were generated with no use of statistics at all, and the other two have not an algorithmic form, a systematic comparison according to the scheme adopted above for the QSARs is not feasible. A narrative presentation of our analyses is adopted.

To display the results of each analysis, a Receiver Operating Characteristics (ROC) graph is used. It reports true positive rate (sensitivity) on the Y-axis, and false positive rate (1 - specificity) on the X-axis. In a ROC graph, perfect performance is located at the left upper corner; the diagonal line represents random results (Provost & Fawcett 2001).

3.4.1 Ashby' SAs

Table III lists the Ashby' SAs (Ref. 62). Figure 5 displays the result of the application of the Ashby' SAs to the 3 databases. The figure indicates that the agreement between Ashby' SAs, and mutagenicity and carcinogenicity respectively, is very similar for CPDB and ISSCAN. The agreement with the Toxnet mutagenicity data is slightly lower than that with the CPDB and ISSCAN mutagenicity data, but still a cluster of "mutagenicity predictions" is clearly apparent. In quantitative terms, the agreement (accuracy) between Ashby's SAs and *Salmonella* mutagenicity

outcomes is 0.73 to 0.79, whereas the agreement with the rodent carcinogenicity outcomes is about



10% lower, i.e., 0.62 to 0.65.

Figure 5. The agreement between the Ashby's SAs, and the mutagenicity and carcinogenicity calls in various databases is shown. (Circles: mutagenicity databases; Triangles: carcinogenicity databases).

There is wide evidence demonstrating that among the short-term mutagenicity assays, the Ames test has the highest correlation with, and predictive ability for rodent carcinogenicity (Benigni 1995; Zeiger, Haseman, Shelby, Margolin & Tennant 1990) (Zeiger 1994). Thus, the Ashby's SAs and the Ames test were compared for their ability to predict the carcinogenicity data in the CPDB and ISSCAN databases. Figure 6 shows that the agreement between Ames test and rodent carcinogenicity is of the same order of magnitude of that between SAs and rodent carcinogenicity.

Table III: Ashby' Structural Alerts

1. Alkyl esters of either phosphonic or sulphonic acids;
2. Aromatic nitro groups;
3. Aromatic azo groups;
4. Aromatic rings N-oxides;
5. Aromatic mono- and dialkylamino groups;
6. Alkyl hydrazines;
7. Alkyl aldehydes;
8. N-methylol derivatives;
9. Monolakenes;
10. N and S, β -haloethyl;
11. N-chloroamines;
12. Propiolactones and propiosultones;
13. Aromatic and aliphatic aziridinyll derivatives;
14. Aromatic and aliphatic substituted primary alkyl halides;
15. Derivatives of urethane (carbarnates);
16. Alkyl N-nitrosoamines;
17. Aromatic amines (including their N-hydroxy derivatives and the derived esters);
18. Aliphatic and aromatic epoxides;
19. Polycyclic aromatic hydrocarbons.

It should be remarked that the pattern of relationships among animal (rodent carcinogenicity), *in vitro* (*Salmonella*) and theoretical (SAs) models displayed in Figure 6 is not trivial. It confirms, in quantitative terms, that the SAs and the *Salmonella* assay are two different representations of the same mechanistic knowledge on chemical carcinogenicity derived from the seminal work of the

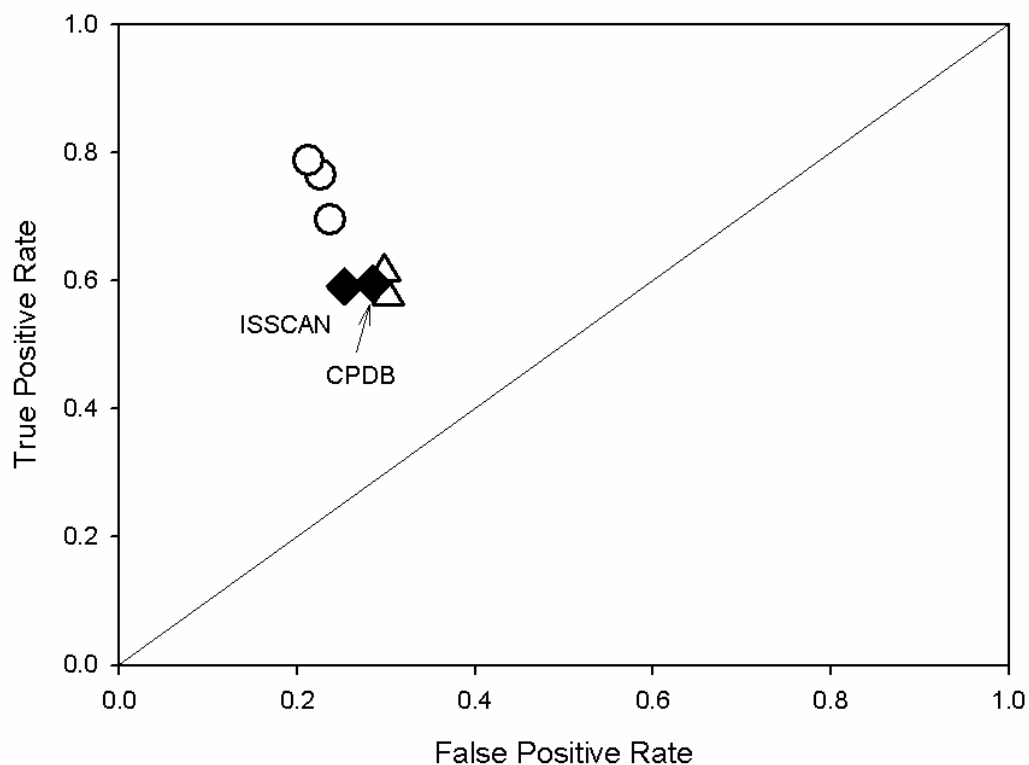


Figure 6. Together with the information reported in Figure 4 (empty symbols), the ROC graph shows the agreement between the *Salmonella* mutagenicity results and rodent carcinogenicity in the CPDB and ISSCAN databases (filled rhombuses).

Millers (Miller & Miller 1981a; Miller & Miller 1981b) (and subsequent investigations). Thus, they have a similar degree of correlation with rodent carcinogenicity. At the same time, even though stemming from chemical carcinogenicity mechanistic knowledge, the SAs agree with *Salmonella* data better than with carcinogenicity data. In fact, they represent the DNA reactive (genotoxic) mechanisms of carcinogenicity, which were the main subject of study of the Millers and which were also the basis for the construction of the *Salmonella typhimurium* strains used for the Ames test (Ames 1984). This explains the higher correlation of the SAs with *Salmonella* than with carcinogenicity.

Figure 6 also points to what is missing in the Ashby' SAs to obtain a higher agreement with rodent carcinogenicity along the Sensitivity dimension (Y-axis in the ROC graphs): the knowledge

on nongenotoxic carcinogens. Regarding the limitations of the SAs in Specificity (X-axis), it can be hypothesized that the knowledge coded into the SAs is a poor representation of the Absorption, Distribution, Metabolism and Excretion (ADME) component.

3.4.2 Bailey' SAs

Table IV lists the Bailey' SAs (Ref. 65); Figure 7 displays the agreement between this set of SAs, and the *Salmonella* and rodent carcinogenicity data. For a comparison, the Ashby' SAs are also included. Whereas the two sets of SAs have similar sensitivity, they are different in terms of specificity. The Bailey et al. SAs give rise to more false positive responses in respect to Ashby' SAs. This lower specificity may be explained with a more conservative approach taken in the regulatory context of the generation of the SAs list. In other terms, the Bailey' SAs are a more conservative version of the Ashby' SAs.

Table IV: Bailey' Structural Alerts

- | |
|--|
| <ol style="list-style-type: none"> 1. Primary and secondary aromatic amines (with methyl or ethyl, or activated methyl or ethyl, substituents) 2. Tertiary aromatic amines (with methyl or ethyl substituents) 3. Secondary aromatic acetamides and formamides 4. Nitroarenes 5. Nitrosoarenes 6. Arylhydroxylamines 7. N-nitroso-N-dialkylamines 8. N-nitroso-N-alkylamides 9. N-nitroso-N-alkylureas 10. N-nitroso-N-alkylcarbamates 11. N-nitrosos-N-alkylnitriles |
|--|

12. N-nitroso-N-hydroxylamines
13. Hydrazines
14. Azoxy alkane
15. Aliphatic halides
16. Benzylic halides
17. Oxiranes and aziridines
18. Propiolactones
19. Alkyl esters of sulfonic and sulphuric acids (with methyl or ethyl substituents)
20. Alkyl esters of phosphonic and phosphoric acids (with methyl or ethyl substituents)
21. Mixed alkyl esters of phosphoric with methyl or ethyl substituents)
22. Haloethylamines
23. Haloalkylethers (ethyl and methyl)
24. α -Halocarbonyl or α -halohydroxy
25. Haloamines
26. α,β -unsaturated carbonyls (aldehyde, ketone, ester, or amide group)
27. Allylic halides and alkoxides (Cl, Br or I)
28. Halogenated methanes
29. Vinyl halides (Cl, Br or I)
30. Polycyclic aromatic hydrocarbons
31. Isocyanate
32. Isothiocyanate
33. Azoarenes (sulfonic group on both rings non-alerting)

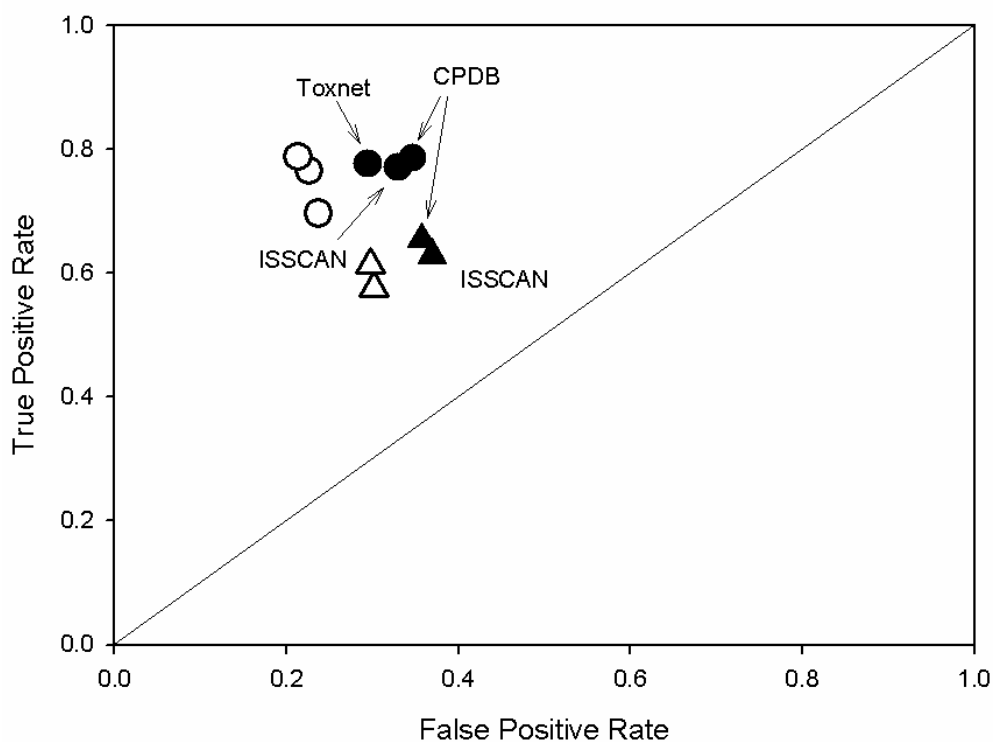


Figure 7. The agreement between the Bailey' SAs, and the mutagenicity and carcinogenicity calls in various databases is shown. For a comparison, also the performance of the Ashby' SAs is shown. (Circles: mutagenicity databases; Triangles: carcinogenicity databases; Filled symbols: Bailey' SAs; Empty symbols: Ashby' SAs).

3.4.3 Kazius et al., 2005, SAs

Table V lists the SAs derived by Kazius et al., 2005 (Ref. 63). Figure 8 reports the application of this set of SAs to the probes. Overall, this set of SAs did not perform much differently from Ashby's SAs (a small increase in Sensitivity is balanced by a small decrease in Specificity). An expected result is that the Kazius 2005 SAs performed better than the Ashby' SAs on the genotoxicity database (Toxnet), which was used as training set.

Table V: Kazius' Structural Alerts (2005)

1. Specific aromatic nitro
2. Specific aromatic amine
3. Aromatic nitroso
4. Alkyl nitrite
5. Nitrosamine
6. Epoxide
7. Aziridine
8. Azide
9. Diazo
10. Triazene
11. Aromatic azo
12. Unsubstituted heteroatom-bonded heteroatom
13. Aromatic hydroxylamine
14. Aliphatic halide
15. Carboxylic acid halide
16. Nitrogen or sulfur mustard
17. Bay-region in polycyclic aromatic hydrocarbons
18. K-region in polycyclic aromatic hydrocarbons
19. Polycyclic aromatic system
20. Sulfonate-bonded carbon (alkyl alkane sulfonate or dialkyl sulfate)
21. Aliphatic <i>N</i> -nitro
22. α,β -unsaturated aldehyde (including R-carbonyl aldehyde)
23. Diazonium
24. β -propiolactone
25. α,β -unsaturated alkoxy group

26. 1-aryl-2-monoalkyl hydrazine
27. Aromatic methylamine
28. Ester derivative of aromatic hydroxylamine
29. Polycyclic planar system

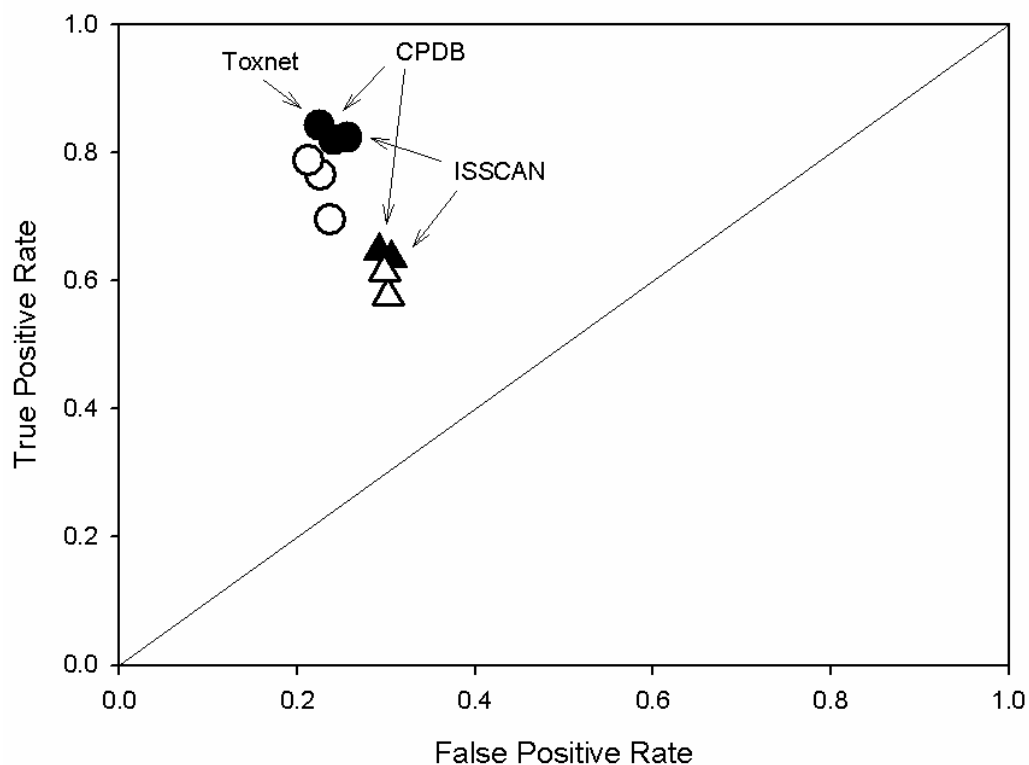


Figure 8. The agreement between the Kazius' SAs (first set), and the mutagenicity and carcinogenicity calls in various databases is shown. For a comparison, also the performance of the Ashby' SAs is shown.

(Circles: mutagenicity databases; Triangles: carcinogenicity databases;
Filled symbols: Kazius' SAs; Empty symbols: Ashby' SAs).

3.4.4 Kazius et al., 2006, SAs

The fourth set of SAs (Table VI) considered was generated by Kazius et al., 2006, (Ref. 64) as an exercise of application of machine learning methods, and has to be judged within this perspective. Figure 9 shows that this set of SAs was inferior to the Ashby's SAs as agreement with *Salmonella* data, whereas had a higher specificity but somewhat lower sensitivity for the carcinogens.

Table VI: Kazius' Structural Alerts (2006)

- | |
|--|
| <ol style="list-style-type: none">1. Highly branched substructure, composed by 11 planar atoms connected with planar bonds;2. Nitrogen atom connected through a double bond to a nitrogen or an oxygen atom;3. Aliphatic epoxides and aziridines;4. Aliphatic halogen (chlorine, bromine, and iodine);5. Aromatic primary amine;6. Heteroatom (N, O)-bonded heteroatom (NH, OH) substructure. |
|--|

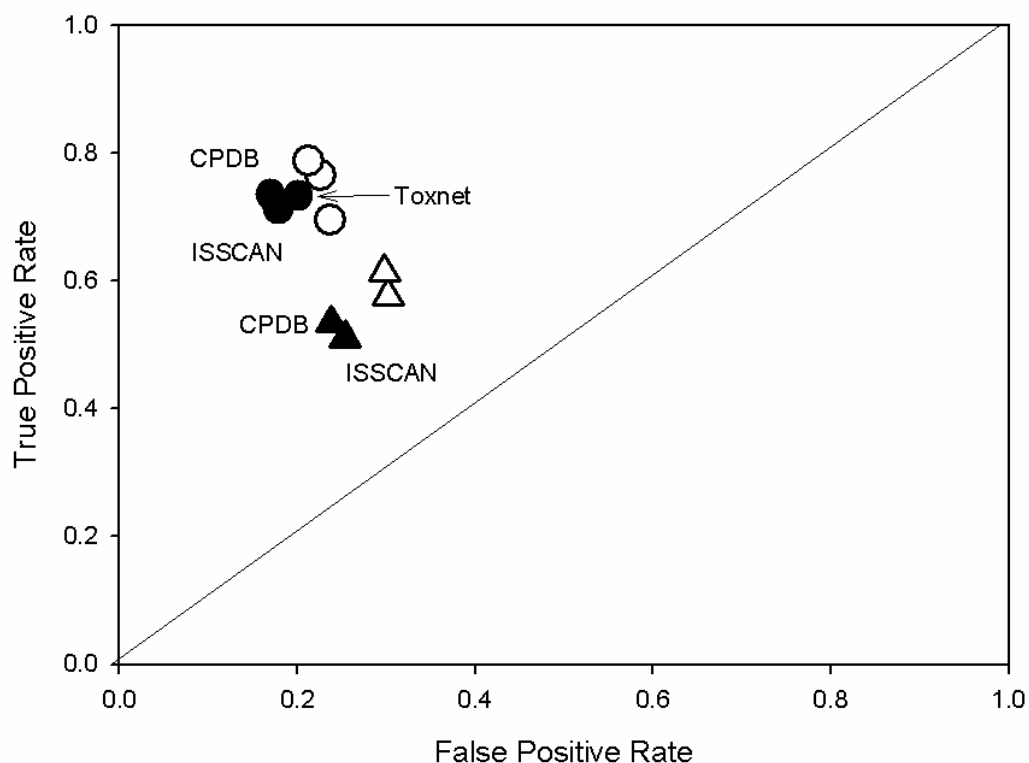


Figure 9. The agreement between the Kazius' SAs (second set), and the mutagenicity and carcinogenicity calls in various databases is shown. For a comparison, also the performance of the Ashby' SAs is shown.

(Circles: mutagenicity databases; Triangles: carcinogenicity databases;
Filled symbols: Kazius' SAs; Empty symbols: Ashby' SAs).

3.4.5 Non-general databases as probes

Overall, the four sets of SAs are not remarkably different from each other in terms of agreement with *Salmonella* mutagenicity and rodent carcinogenicity. What is interesting is the fact that they all show a similar pattern of correlation with the two endpoints. In turn, *Salmonella* data demonstrated a relationship with carcinogenicity data similar to that of the SAs.

Figures 5 to 9 report an estimation of the “average” ability of different sets of SAs to identify mutagens / carcinogens in large sets of data, including the majority of the chemicals with environmental relevance tested so far. It is interesting to investigate also how this average value is

modulated when the SAs models are applied to subsets of chemicals selected according to very specific criteria. Here, two extreme cases have been investigated. The first one regards a group of congeneric chemicals (aromatic amines) and the second one regards a group of pharmaceuticals, with very diverse chemical structures.

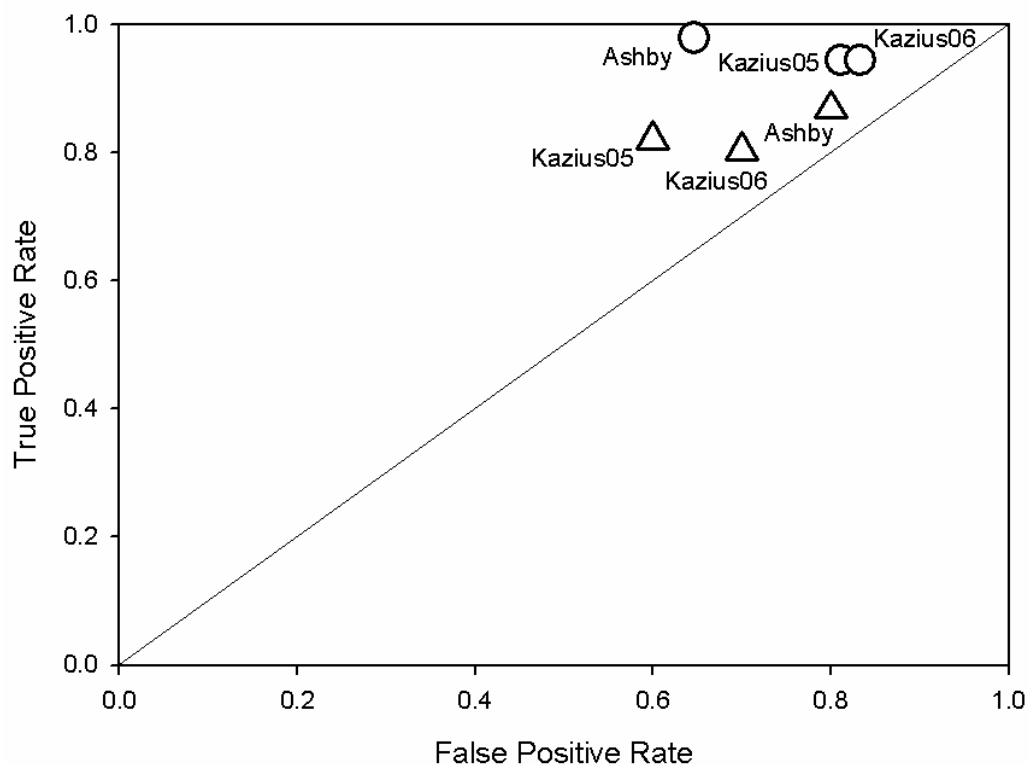


Figure 10. The agreement between the Ashby's and Kazius' SAs (first (05) and second (06) set), and the mutagenicity and carcinogenicity calls in a database of aromatic amines is shown. (Circles: mutagenicity; Triangles: carcinogenicity).

The aromatic amines are the chemical class with the largest amount of available experimental data. The mutagenicity and the carcinogenicity data used for this analysis were retrieved from the compilations of (Debnath, Debnath, Shusterman & Hansch 1992) and (Franke, Gruska, Giuliani & Benigni 2001). Figure 10 shows the application of different sets of SAs to the aromatic amines. Whereas the sensitivity is very high, the specificity is very low. The sensitivity is high for trivial

reasons, because all the chemicals contain at least one alert, i.e. an amino group. Regarding the specificity, it appears that no set of SAs is able to discriminate efficiently between the aromatic amines that are actually toxic, and those whose potential is not expressed in the experimental system. Thus, the application in Figure 10 indicates that the SAs considered are poorly suitable to express the gradation of effects that different molecular environments exert on the potentially DNA reactive moiety.

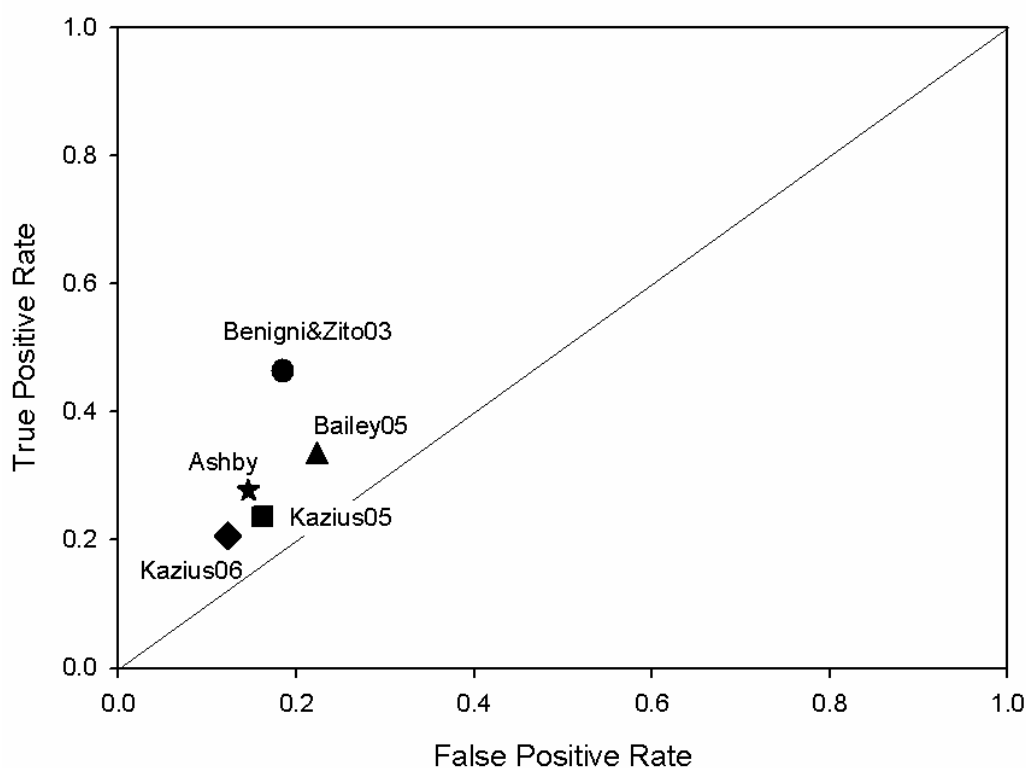


Figure 11. The ROC graph displays the agreement between various sets of SAs and the rodent carcinogenicity calls in a database of pharmaceuticals. For a comparison, the predictions by a human expert (Romano Zito) (Benigni & Zito 2003) are reported.

A second comparison of SAs on a specially selected subset of chemicals was performed on a database of pharmaceuticals with carcinogenicity data (Figure 11). This dataset was previously used as test set in a Predictive Toxicology Challenge (PTC) on rodent carcinogenicity (Benigni & Giuliani 2003; Helma & Kramer 2003). In the PTC exercise, the carcinogenicity of the

pharmaceuticals test set was predicted with a range of machine learning algorithms, which had been previously trained on a set of industrial chemicals (training set). Overall, the result of the PTC exercise was quite deceiving, one advocated reason being that the training (industrial) and test (pharmaceutical) sets were structurally quite dissimilar.

Figure 11 supports the PTC conclusions. The sets of SAs performed quite poorly on the pharmaceuticals, mainly due to low sensitivity: many pharmaceutical carcinogens remained unnoticed because appropriate SAs are not available. Since many carcinogens in the pharmaceuticals set are negative in the mutagenicity assays, it can be assumed that they act through nongenotoxic mechanisms. This result stresses the need for developing lists of SAs for the nongenotoxic mechanisms of chemical carcinogenicity (Woo 2003). For the sake of comparison, Figure 11 reports the results of the prediction of the carcinogenicity of the PTC pharmaceuticals by a human expert (Romano Zito) (Benigni & Zito 2003). It appears that the human expert outperformed all the sets of SAs in sensitivity, thus indicating the existence of an additional body of knowledge that can be potentially transformed into formalized rules (e.g., SAs).

4 Conclusions

This evaluation of the non-commercial (Q)SARs for mutagenicity and carcinogenicity consisted of a preliminary survey (Phase I), and then of a more detailed analysis of short listed models (Phase II). In Phase I, the models were collected from the literature, and then assessed according to the OECD principles –based on the information provided by the authors-; thus Phase I provided the support for short listing a number of promising models (Table I), that were analyzed more in depth in Phase II. In Phase II, the information provided by the authors was completed and complemented with a series of analyses aimed at generating an overall profile of each of the short listed models. This included statistical analyses of the models, generation of external test sets for the assessment of external predictivity, generation of new models for activity. Given their different natures, the assessment scheme for the QSAR models for congeneric sets of chemicals was different from that adopted for the global, or non-local models.

4.1 QSARs for congeneric sets of chemicals

The literature contains QSARs for most of the top ranking classes of the EU HPV list (Table II), with the notable exception of the halogenated aliphatics. For this important chemical class only models for the genotoxic activity in *Aspergillus nidulans* (aneuploidy) exist (Refs. 51 and 52 in Appendix 1), but this assay has no regulatory recognition (thus, Refs. 51 and 52 were not included in Phase II of this work).

The congeneric QSARs almost exclusively aim at modeling *Salmonella* mutagenicity and rodent carcinogenicity, which are crucial toxicological endpoints in the regulatory context. However, the lack of models for *in vivo* genotoxicity should be remarked. According to an assessment carried out by the European Chemicals Bureau (ECB), the *in vivo* mutagenicity studies, shortly followed by carcinogenicity, are posing the highest demand for test-related recourses (Pedersen, de Bruijn,

Munn & Van Leeuwen 2003; Van der Jagt, Munn, Torslov & de Bruijn 2004). In particular, the *in vivo* micronucleus test is widely used for regulatory purposes as follow-up to bacterial mutagenicity, and requires the sacrifice of large numbers of animals. A QSAR alternative is desirable.

Many available QSARs model the potency of active chemicals (mutagens or carcinogens) only. However, it is recognized that in the field of mutagenicity and carcinogenicity the difference between actives and inactives, and the gradation of potency of the actives may depend on different chemical properties, so they should be modeled separately (Benigni, Andreoli & Giuliani 1994; Franke, Gruska, Giuliani & Benigni 2001). When possible (i.e., mutagenic activity of aromatic amines in two Salmonella strains), we filled the gaps by generating new models specifically for this project (QSAR5 and QSAR6).

Overall the short listed models -either reported by other authors or generated by us- can be interpreted mechanistically, and they agree with, and/or support the available scientific knowledge.

Regarding statistics (e.g., fitting parameters, cross-validation, etc...), the data provided by the authors together with the measures calculated by us pointed out that –overall- the short listed models are of good quality (except maybe some low values in cross-validation for QSAR3, QSAR12, and QSAR13. However, this did not influence the ability of predicting external test sets – see below-).

A crucial point is that of “validation”. Whereas it is generally accepted that the gold standard is to test the model on a set of chemicals not used for the derivation of the model, in practice many investigators use different statistical procedures to generate artificial test sets (e.g., by splitting the sample of chemicals into two sets, and regarding one as training and the other one as test set). However, it is our opinion that all these internal validation procedures only generate different statistical descriptions of the same original data subjected to modeling. A crucial and necessary further assessment is to subject the model to a real external validation test: the activity of congeneric chemicals -not considered in any way at the moment of the generation of the model- are

predicted, and the predicted and experimental activities are compared. Thus, we searched the literature for external data sets not considered by the authors of the models, and whenever the data were available, we performed external prediction exercises (for QSAR1 to QSAR10. It has not been possible for QSAR11, whereas for QSAR13 the external predictivity was assessed by the authors).

In the selection of the external sets, the constraint for a test set to belong to the Applicability Domain of the training set was taken into account by considering: a) the types of structures to which the model applies. This was assessed subjectively by us according to our expert knowledge, by checking, among others, the absence of reactive groups different from those that characterize the chemical class under study; b) the ranges of descriptors values of the two sets; c) a mathematical transform of structural similarity indices (ranges of PC scores). Criterion a) was considered crucial, and Criteria b) and c) were considered as confirmatory. A preliminary assessment showed that the external test sets used by us were by and large within the applicability domains of the models (i.e., training sets), and that minor deviations did not affect the goodness of prediction. For example, in QSAR1 there is only one chemical in the test set that has a logP value outside the range of the training set, but for all the other parameters (HOMO, LUMO, and similarity indices) the training and test sets overlap.

The results of the external prediction tests are reported in the individual sections. These results, viewed together with the various statistical measures on the training sets, provide very interesting evidence. The following is the description of our results.

Table VII summarizes the external prediction outcomes for regression based models (i.e., QSAR models for potency), and Table IX summarizes the outcomes for discriminating models (i.e., QSAR models for activity). The two tables report also parameters for goodness of fit and different internal validations of the training set. In addition, Table VIII shows the correlation coefficients among the parameters in Table VII, and Table X shows the correlation coefficients those in Table IX.

Inspection of Table VII indicates that the goodness of fit in the training set (correlation coefficient, *r_{tra}*) is always considerably better than the goodness of prediction (*r_{te}*) for the test set.

rte is the correlation coefficient between predicted and experimental potency of the test set. It also appears that the internal validation measures (*q2* and *q2_10*) are lower than the back-fitting of the model (*rtra*), but still considerably higher than the external prediction *rte*. This suggests that the internal validation measures are bad predictors of the performance with external test sets.

Table VII : Regression-based models for Potency: fit and predictivity measures

QSAR	System	<i>rtra</i>	<i>q2</i>	<i>q2_10</i>	<i>lever</i>	<i>rte</i>	<i>accte</i>
qsar1	TA98	.90	.78	.71	.06	.41	.36
qsar2	TA100	.88	.74	.66	.06	.68	.57
qsar3	mouse	.91	.58	.0	.25	.56	.58
qsar4	rat	.93	.81	.79	.15	.48	.71
qsar9	TA98	.90	.89	.80	.04	-.23	.43
qsar10	TA100	.88	.77	.73	.05	.36	.32

An alternative way of measuring the prediction performance for regression based models is to calculate its accuracy as percentage of test chemicals correctly predicted within one log unit of activity (*accte*). When expressed as *accte*, the prediction performance is a more robust estimate than when expressed as *rte* (see for example QSAR9, which shows a negative correlation between predicted and experimental potency values (*rte*), whereas the percentage of chemicals correctly predicted within one log unit (*accte*) is 0.43). This is understandable, since high *rte* values require exact point estimates, whereas high *accte* requires correct estimates of intervals; the latter is a less stringent criteria and, in addition, is closer to the regulatory needs. Overall, the QSAR external predictions for the potency of congeneric chemicals are 30 to 70 % correct (as seen as *accte*).

Table VIII indicates that *rtra* is correlated with *accte* but not with *rte*, thus indirectly confirming that *accte* is a performance index more robust than *rte*.

Regarding the internal validation indices $q2$, $q2_10$ and mean Leverage ($lever$), Table VIII shows that $q2$ and $q2_10$ (Leave-10%-Out crossvalidation) are negatively correlated with both rte and $accte$. In addition the mean Leverage ($lever$), whose high values are supposed to indicate “bad” models with uneven influence of individual data points on the models themselves, are positively correlated with both the external validation indices rte and $accte$. All these results are contrary to what one could expect.

Table VIII: Regression-based models for Potency: Correlation Coefficients

	$rtra$	$q2$	$q2_10$	$lever$	rte	$accte$
$rtra$	1.00	0.09	-0.02	0.52	-0.07	0.60
$q2$		1.00	0.93	-0.77	-0.69	-0.25
$q2_10$			1.00	-0.84	-0.39	-0.25
$lever$				1.00	0.43	0.62
rte					1.00	0.41
$accte$						1.00

The results of external validation for the discriminant models (activity) are in Table IX. As in the case of regression based models, the overall accuracy in the training set ($acctr$) is systematically higher than that attained in the external test set ($accte$). However, the external prediction performance is 63 to 100 % accurate, considerably higher than that of the regression models for potency (30 to 70 %). This confirms the evidence that predicting intervals is more reliable than predicting individual data points.

Table IX: Discriminant models for Activity: fit and predictivity measures

QSAR	System	<i>sqcc</i>	<i>acctr</i>	<i>acc10</i>	<i>accte</i>
qsar7	rodent	0.38	0.88	0.75	0.67
qsar8	rodent	0.50	0.94	0.78	0.70
qsar5	TA98	0.31	0.83	0.83	0.63
qsar6	TA100	0.39	0.79	0.65	0.69
qsar13	TA100	0.61	1.0	0.85	1.0

More information is reported in Table X, which shows that the accuracy in the training set (*acctr*) is a good predictor of external predictivity (*accte*). An even better predictor of external predictivity is the Squared Canonical Correlation (*sqcc*) of the model.

Table X also shows that, as in the case of regression based models, the internal validation index (here *acc10*: cross-validation Leave-10%-Out) is a mediocre indicator of external predictivity.

Table X: Discriminant models for Activity: correlation coefficients

	<i>sqcc</i>	<i>acctr</i>	<i>acc10</i>	<i>accte</i>
<i>sqcc</i>	1.00	0.89	0.37	0.90
<i>acctr</i>		1.00	0.68	0.78
<i>acc10</i>			1.00	0.45
<i>accte</i>				1.00

Whereas generally the importance of crossvalidation is overestimated in the QSAR literature, a few authors have pointed to the limitations of crossvalidation as an assessment of the “goodness” of a model (e.g., (Golbraikh & Tropsha 2002) (Kubinyi 2005)). This report, based on a number of real

case studies and not only on computer simulations, adds new supporting evidence, and expands the reach of the previous observations on the limits of internal validation. Thus, complementary roles can be envisaged for internal and external validation procedures. The internal validation procedures are useful tools in the phase of the model construction (statistical consistency), whereas external validation assesses the confidence one can have in the predictions of the model itself.

4.2 SARs for non-congeneric data sets

Among the non-local, or global approaches for non-congeneric data sets, four models based on the use of Structural Alerts (SA) were short listed and investigated in more depth (Table I). These models were selected because of the overriding importance of the SAs as a basis for evaluations by human experts in the regulatory context, and for implementations into computer programs. In order to assess the four sets, it was necessary to preliminarily build in our laboratory the capacity to implement the four sets into one computer platform: this allowed us to compare in a systematic way the four sets on the same probes (i.e., databases of mutagens and carcinogens).

Overall, the four sets did not differ to a large extent in their performance. In the “general” databases the SAs appear to agree around 65% with rodent carcinogenicity data, and 75% with Salmonella mutagenicity data. These figures can be considered as “general” measures of the predictive ability of the SAs for the “known” universe of chemicals associated with experimental data (which is not the entire universe of all possible chemicals).

On the other hand, SAs based models do not seem to work equally efficiently in the discrimination between active and inactive chemicals within individual chemical classes (see the exercise on the class of the aromatic amines). This because of the lack of detailed sub-rules describing how each alert is modulated by the different molecular environments. This kind of more refined modeling is the purpose of the QSARs for individual chemical classes (which, however, exist only for a limited number of them).

It should be remarked that the evolving societal have a strong influence on the database of “important” chemicals and, consequently, on the predictive ability of the SAs. Table XI shows the distribution of the SAs (our unpublished results) among a set of chemicals bioassayed by the US National Toxicology Program (NTP) in the recent years (<http://ntp.niehs.nih.gov/>). It appears that only a few of them have recognizable SAs, and that several rodent carcinogens do not posses SAs at all. For example, Ashby’ SAs are in 3 out 25 carcinogens, and Bailey’ SAs are in 6 out 25 carcinogens. In addition, several carcinogens are non-mutagenic (only 8 out 22 carcinogens are *Salmonella* mutagens): hence, they are putatively nongenotoxic. The progressive decline of the proportion of genotoxic carcinogens during the years is a positive effect of the increased knowledge on the genotoxic mechanisms of carcinogenesis; this has permitted a reduction of the number of new chemicals with known SAs that have been put into the market. However, this temporal trend also diminishes our ability to recognize carcinogens through the established lists of SAs, and sets up a different scenario to be faced by predictive systems. In this new scenario, the need of formulating models for nongenotoxic carcinogens is a priority.

In addition to the above area, there are two more areas where the expansion of the SAs is desirable. First, the lack of SAs specific for *in vivo* mutagenicity assays, e.g., micronucleus, that rely on the sacrifice of considerable numbers of animals, should be remarked. In fact, it is known that the classical genotoxicity assays (e.g., *Salmonella*) are not overlapping –in terms of endpoint and of chemicals to which are sensitive- with the *in vivo* assays (Benigni 1995). Thus, this area deserves more research.

Second, our practical experience with the implementation of the SAs has indicated that also here there is space for further technical improvement. Beside the scientific principles, we think that a systematic effort to combine the best of the various sets of SAs and to expand the sub-rules is necessary.

NTP	ChemName	CAS	Ashby	Bailey	Kazius05	Kazius06	Canc	Sty
TR500	Naphtalene	91-20-3	no	no	no	no	+	-
TR501	4,4[-Dichlorodiphenyl sulphone]	80-07-9	no	no	no	no	-	-
TR502/503	Chloral Hydrate	302-17-0	no	no	yes	yes	+	+
TR504	o-Nitrotoluene	88-72-2	yes	yes	yes	yes	+	-
TR505	Citral	5392-40-5	yes	yes	N.A.	N.A.	?	-
TR506	Acrylonitrile	107-13-1	no	no	N.A.	N.A.	+	+
TR507	Vanadium Pentoxide	1314-62-1	no	no	no	no	+	-
TR508	Riddelliine	23246-96-0	no	yes	no	no	+	+
TR509	2,4-Hexadienal	142-83-6	yes	yes	N.A.	N.A.	+	+
TR511	Dipropylene Glycol	25265-71-8	no	no	N.A.	N.A.	-	-
TR512	Elmiron	37319-17-8	no	no	yes	no	+	-
TR513	Decalin	91-17-8	no	no	N.A.	N.A.	+	-
TR514	trans-Cinnamaldehyde	14371-10-9	yes	yes	N.A.	N.A.	-	-
TR515	Propylene Glycol Mono-t-Butyl Ether	57018-52-7	no	no	no	no	+	+
TR516	2-Methylimidazole	693-98-1	no	no	N.A.	N.A.	+	-
TR517	Sodium Chlorate	7775-09-9	no	no	no	no	+	-
TR518	Triethanolamine	102-71-6	no	no	N.A.	N.A.	+	-
TR520	3,3',4,4',5-Pentachlorobiphenyl (PCB 126)	57465-28-8	no	no	no	no	+	N.D.
TR521	2,3,7,8-Tetrachlorodibenzo-p-dioxin (TCDD)	1746-01-6	no	no	N.A.	N.A.	+	-
TR522	Transplacental AZT	30516-87-1	no	yes	N.A.	N.A.	+	+
TR523	Diisopropylcarbodiimide	693-13-0	no	no	N.A.	N.A.	-	-
TR525	2,3,4,7,8-Pentachlorodibenzofuran (PeCDF)	57117-31-4	no	no	yes	yes	+	N.D.
TR527a	Malachite Green Chloride	569-64-2	yes	yes	no	no	?	+
TR527b	Leucomalachite Green	129-73-7	yes	yes	N.A.	N.A.	?	-
TR529	2,2',4,4',5,5'-Hexachlorobiphenyl (PCB 153)	35065-27-1	no	no	no	no	?	N.D.
TR532	Bromodichloromethane	75-27-4	no	no	yes	yes	+	-
TR533	Benzophenone	119-61-9	no	no	N.A.	N.A.	+	-
TR534	Divinylbenzene-HP	1321-74-0	no	no	N.A.	N.A.	?	-
TR535	4-Methylimidazole	822-36-6	no	no	no	no	+	-
TR537	Dibromoacetic Acid	631-64-1	no	no	N.A.	N.A.	+	+
TR538	Methyl Isobutyl Ketone	108-10-1	no	no	N.A.	N.A.	+	-
TR540	Methylene Blue Trihydrate	7220-79-3	yes	yes	N.A.	N.A.	+	+
TR543	α -Methylstyrene	98-83-9	no	no	N.A.	N.A.	+	-
TR545	Genistein	446-72-0	no	yes	no	no	+	N.D.

Table XI. Agreement between four sets of SAs and carcinogenicity and mutagenicity calls in the chemicals most recently bioassayed by the US National Toxicology Program. The chemicals are identified by the Technical Report (TR) number, chemical name and CAS number. Yes / no indicates the presence of SAs (according to the different lists). N.A. (=Not Applicable) indicates that the chemical is in the training set of the model, thus the prediction is not considered. N.D. = No Data. +, -, ? indicates positive, negative, equivocal response to the carcinogenicity (Canc) and *Salmonella* mutagenicity (Sty) assays.

A final comment on the issue of Applicability Domain is necessary. We did not attempt to define the AD of the four sets of SAs in the same systematic way as we did for the QSARs. Rigorously speaking, the AD of a certain SA is only the set of chemicals that contain that SA. An alternative (practical) approach is to consider as AD of a set of SAs, the chemicals (or some representation of them) that were used to derive that set of SAs. But in many cases such a set of chemicals cannot be defined, e.g., for the Ashby' or Bailey SAs, that originate from a complex body of mechanistic knowledge with different origins. To make things more contradictory, when we tested the four sets of SAs on a database of pharmaceuticals (the PTC chemicals) we measured the similarity (Tanimoto coefficients) of the PTC chemicals in respect to the chemicals in "general" databases (e.g., ISSCAN or CPDB). In that case, we found that the PTC chemicals were not different from some clusters of ISSCAN/CPDB chemicals (results not shown). In spite of this similarity, the prediction ability of the SAs was very different in the general databases and in the PTC chemicals (see, for example Fig. 5 *versus* Fig. 11). Even though it can be argued that the adoption of different similarity measures may generate different similarity patterns, our previous research has shown that the similarity measure used may be critical at the small scale of congeneric chemicals, but for large and diverse data sets this choice usually has minor consequences on the resulting similarity values (Benigni, Gallo, Giorgi & Giuliani 1999). Overall, our opinion is that more research is still needed on the subject of the definition of AD issue for SAs.

4.3 Final considerations

A general indication of this study, valid for both congeneric and noncongeneric models, is that there is uncertainty associated with (Q)SARs. Thus a prediction cannot be taken at face value, and the level of uncertainty has to be considered when using (Q)SAR in a regulatory context. However, (Q)SARs are not meant to be black-box machines for predictions, but have a much larger reach and scope. Here applies the same reflection made by Rainer Franke regarding the search for new drugs:

“As the drug discovery process is of a very complex nature, effective drug design requires an entire spectrum of techniques in which QSAR methods still play an important role. ... The real power of drug design methods is to extract and synthesize information from data to obtain hypotheses that can be put to experimental test. No dramatic overnight discoveries of wonder drug will result, but an increase in the chance of success due to indications of promising directions is a realistic expectation....” (Franke & Gruska 2003). Equally, a regulatory process often consists of a complex assessment that requires the combination of different types of evidence, among which there are (Q)SARs. Using only non-testing methods, the larger the evidence from QSARs (several different models, if available) and other approaches (e.g. chemical categories, read across) the higher the confidence in the prediction.

Regarding the SAs, their main role is that of preliminary, or large-scale screenings. This has been brilliantly demonstrated by the selection process for chemicals to be bio-assayed by the US National Toxicology Program. Two thirds of the chemicals selected for the bioassay were “suspect” chemicals: this selection was operated by human experts largely based on the recognition of SAs. Another one third was selected only on production/exposure considerations. As a matter of fact, the proportion of carcinogens among the “suspected” chemicals was almost ten times higher than that relative to the chemicals selected only on production/exposure considerations (Fung, Barrett & Huff 1995). Thus, the recognition of SAs was a powerful tool to enrich considerably the target of the priority setting. It can be anticipated that a similarly successful role can be played by the SAs to support the process of grouping chemicals for read-across and category formation.

In comparison with the SAs and qualitative SARs, the QSARs for individual chemical classes – when available- can be used with higher confidence. Evidence produced in this work has indicated that predictions of intervals (e.g., negative / positive) are more reliable than predictions of exact data points (e.g., potency).

Another general indication of this study is that a better balance between statistics on one side, and mechanistic knowledge and (Q)SAR know how on the other side is needed. As a matter of fact, one

result of our analysis on the QSARs for congeneric sets shows that some of the commonly accepted measures of internal validation (often generalized as “validation” *tout court*) is that they are very poor predictors of the ability of models to perform well with external data sets. Deriving and evaluating (Q)SAR models should not only rest on statistical criteria but should also properly take into account mechanistic aspects. This is crucial, since mechanistic interpretation is a very powerful tool for deciding about the validity of a (Q)SAR. In addition, it provides a ground for interaction and dialogue between model developers, and toxicologists and regulators, and permits the integration of the (Q)SAR results into a wider regulatory framework, where different types of evidence and data concur or complement each other as a basis for making decisions and taking actions.

An example can clarify further the advantages of the practice of using mechanistic (Q)SAR in the regulatory process. For example, recognizing that a certain chemical contains a SA will suggest to the regulator a possible mechanism of action. In addition, a QSAR (e.g., QSAR5, for identifying the mutagenic aromatic amines) will indicate also if the chemical is below or above a given reactivity threshold, and if the potential reactivity center is sterically hindered or the metabolic system has free access to it. Thus the information provided by (Q)SARs, even in the presence of the inherent uncertainty linked to the (Q)SARs themselves, will expand the knowledge on the chemical and will help the regulator to put into context other available evidence.

5 References

Ames BN. 1984. The detection of environmental mutagens and potential carcinogens. *Cancer* 53:2030-40.

Ashby J. 1985. Fundamental structural alerts to potential carcinogenicity or noncarcinogenicity. *Environ Mutagen* 7:919-21.

Ashby J, Tennant RW. 1988. Chemical structure, Salmonella mutagenicity and extent of carcinogenicity as indicators of genotoxic carcinogenesis among 222 chemicals tested by the U.S.NCI/NTP. *Mutat Res* 204:17-115.

Benigni R. 1993. Analysis of distance matrices for studying data structures and separating classes. *Quant Struct -Act Relat* 12:397-401.

Benigni R. 1995. Mouse bone marrow micronucleus assay: relationships with in Vitro mutagenicity and rodent carcinogenicity. *Journal of Toxicology and Environmental Health* 45:337-47.

Benigni R. 2005. Structure-activity relationship studies of chemical mutagens and carcinogens: mechanistic investigations and prediction approaches. *Chem Revs* 105:1767-800.

Benigni R, Andreoli C, Giuliani A. 1994. QSAR models for both mutagenic potency and activity: application to nitroarenes and aromatic amines. *Environ Mol Mutagen* 24:208-19.

Benigni R, Bossa C. 2006. Structure-activity models of chemical carcinogens: state of the art, and new directions. *Ann Ist Super Sanità* 42:118-26.

Benigni R, Gallo G, Giorgi F, Giuliani A. 1999. On the equivalence between different descriptions of molecules: value for computational approaches. *J Chem Inf Comput Sci* 39:575-8.

Benigni R, Giuliani A. 2003. Putting the Predictive Toxicology Challenge into perspective: reflections on the results. *Bioinformatics* 19:1194-200.

Benigni R, Richard AM. 1998. Quantitative structure-based modeling applied to characterization and prediction of chemical toxicity. *Methods* 14:264-76.

Benigni R, Zito R. 2003. Designing safer drugs: (Q)SAR-based identification of mutagens and carcinogens. *Curr Top Med Chem* 3:1289-300.

Debnath AK, Debnath G, Shusterman AJ, Hansch C. 1992. A QSAR investigation of the role of hydrophobicity in regulating mutagenicity in the Ames test: 1. Mutagenicity of aromatic and heteroaromatic amines in *Salmonella typhimurium* TA98 and TA100. *Environ Mol Mutagen* 19:37-52.

Franke R, Gruska A. 2003. General introduction to QSAR. In: Benigni R, editor. *Quantitative Structure-Activity Relationship (QSAR) models of mutagens and carcinogens*. Boca Raton: CRC Press; p 1-40.

Franke R, Gruska A, Giuliani A, Benigni R. 2001. Prediction of rodent carcinogenicity of aromatic amines: a quantitative structure-activity relationships model. *Carcinogenesis* 22:1561-71.

Fung VA, Barrett JC, Huff J. 1995. The carcinogenesis bioassay in perspective: application in identifying human cancer hazards. *Environ Health Perspect* 103(7-8):680-3.

Golbraikh A, Tropsha A. 2002. Beware of q²! *J Mol Graph Model* 20:269-76.

Hansch C, Leo A, Hoekman D. 1995. *Exploring QSAR. 2. Hydrophobic, electronic and steric constants*. Washington, DC: ACS American Chemical Society.

Helma C, Kramer S. 2003. A survey of the Predictive Toxicology Challenge 2000-2001. *Bioinformatics* 19:1179-82.

Kazius J, McGuire R, Bursi R. 2005. Derivation and Validation of Toxicophores for Mutagenicity Prediction. *J Med Chem* 48:312-20.

Kubinyi H. 2005. Validation and predictivity of QSAR models. In: Aki-Sener E, Yalcin I, editors. *QSAR and Molecular Modelling in rational design of bioactive molecules. The 15th European Symposium on Quantitative Structure-Activity Relationships and molecular modelling*. Ankara: CADDs; p 30-33.

Miller EC, Miller JA. 1981a. Mechanisms of chemical carcinogenesis. *Cancer* 47:1055-64.

Miller EC, Miller JA. 1981b. Searches for ultimate chemical carcinogens and their reactions with cellular macromolecules. *Cancer* 47:2327-45.

Munro IC, Ford RA, Kennepohl E, Sprenger JG. 1996. Threshold of toxicological concern based on Structure-Activity Relationships. *Drug Metab Rev* 28:209-17.

Pedersen F, de Bruijn J, Munn SJ, Van Leeuwen K. 2003. Assessment of additional testing needs under REACH. Effects of (Q)SARs, risk based testing and voluntary industry initiatives. Ispra: EUR; Report nr JRC report EUR 20863 EN.

Provost F, Fawcett T. 2001. Robust classification for imprecise environment. *Machine Learn J* 42(3):5-11.

Richard AM. 2004. DSSTox web site launch: improving public access to databases for building structure-toxicity prediction models. *Preclinica* 2:103-8.

Richard AM, Williams CR. 2003. Public sources of mutagenicity and carcinogenicity data: use in structure-activity relationship models. In: Benigni R, editor. *Quantitative Structure-Activity Relationship (QSAR) models of mutagens and carcinogens*. Boca Raton: CRC Press; p 145-174.

Sneath PHA. 1983. Distortions of taxonomic structure from incomplete data on a restricted aset of reference strains. J Gen Microbiol 129:1045-73.

Sneath PHA, Johnson R. 1972. The influence on numerical taxonomic similarities of errors in microbiological tests. J Gen Microbiol 72:377-92.

Van der Jagt K, Munn SJ, Torslov J, de Bruijn J. 2004. Alternative approaches can reduce the use of test animals under REACH. Addendum to the Report "Assessment of additional testing needs under REACH. Effects of (Q)SARs, risk based testing and voluntary industry initiatives". Ispra: European Commission Joint Research Centre; Report nr JRC Report EUR 21405 EN.

Verloop A. 1987. The STERIMOL approach to drug design. New York: Marcel Dekker.

Woo YT. 2003. Mechanisms of action of chemical carcinogens, and their role in Structure-Activity Relationships (SAR) analysis and risk assessment. In: Benigni R, editor. Quantitative Structure-Activity Relationship (QSAR) models of mutagens and carcinogens. Boca Raton: CRC Press; p 41-80.

Zeiger E. 1994. Strategies and philosophies of genotoxicity testing: what is the question? Mutation Research 304:309-14.

Zeiger E, Haseman JK, Shelby MD, Margolin BH, Tennant RW. 1990. Evaluation of four in vitro genetic toxicity tests for predicting rodent carcinogenicity: confirmation of earlier results with 41 additional chemicals. Environ Mol Mutagen 16,(suppl.18):1-14.

6 Scheme 1: OECD PRINCIPLES-BASED SCORING SYSTEM FOR THE (Q)SAR MODELS

Defined end-point of regulatory importance: the non-relevant models were eliminated in the preliminary phase (see text); no scores were given;

A) Unambiguous algorithm:

A1) statistical / mathematical procedure clearly described, reproducible based on the information provided;

A2) full details of the training set (A2.1: chemical identity; A2.2: biological data; A2.3: parameters values);

A3) easy access to parameters calculation;

Scores:

A1) yes = 1; no = 2;

A2.1, A2.2, A2.3)) yes = 1; no = 2;

A3) wide availability and acceptance of chemical descriptors = 1;

limited availability of chemical descriptors = 2;

descriptors calculated by only one commercial program = 3;

in-house descriptors = 4;

B) Defined domain of applicability (descriptor space, structure space)

Scores:

descriptor AND structure spaces defined = 1;

descriptor OR structure spaces defined = 2;

none = 3;

C) Measures of

C1) goodness-of-fit (e.g., r);

C2) robustness (e.g., q^2 , leverage, separation into training/test set);

C3) predictivity (external data prediction);

Scores:

C1, C2, C3) yes = 1; no = 2;

D) Mechanistic interpretation

Scores:

Direct, possible = 1; difficult = 2;

E) Others

E1) number of data points

E2) chemical class representation (number) in the EU HPV chemicals.

N.A. = Not Applicable.

7 Appendix 1 List of reviewed papers

1. QSARs for congeneric series

1.1 Aromatic amines

1.1.1 Mutagenic activity

- (1) Trieff NM, Biagi GL, Sadagopa Ramanujam VM, Connor TH, Cantelli-Forti G, Guerra MC, Bunce III H, Legator MS. Aromatic amines and acetamides in Salmonella typhimurium TA98 and TA100: A QSAR study. *Mol Toxicol* 1989; 2:53-65.
- (2) Ford GP, Griffin GR. Relative stabilities of nitrenium ions derived from heterocyclic amine food carcinogens: relations to mutagenicity. *Chem Biol Interact* 1992; 81:19-33.
- (3) Ford GP, Herman PS. Relative stabilities of nitrenium ions derived from polycyclic aromatic amines. Relationship to mutagenicity. *Chem Biol Interact* 1992; 81:1-18.
- (4) Debnath AK, Debnath G, Shusterman AJ, Hansch C. A QSAR investigation of the role of hydrophobicity in regulating mutagenicity in the Ames test: 1. Mutagenicity of aromatic and heteroaromatic amines in Salmonella typhimurium TA98 and TA100. *Environ Mol Mutagen* 1992; 19:37-52.
- (5) Benigni R. QSARs for mutagenicity and carcinogenicity. In: Anonymous, editor. Report from the Expert Group on (Quantitative) Structure-Activity Relationships ((Q)SARs) on the principles for the validation of (Q)SARs. Paris: OECD, 2004: 84-112.
- (6) Basak SC, Grunwald GD. Predicting mutagenicity of chemicals using topological and quantum chemical parameters: a similarity based study. *Chemosphere* 1995; 31:2529-2546.
- (7) Basak SC, Gute BD, Grunwald GD. Assessment of the mutagenicity of aromatic amines from theoretical structural parameters: a hierarchical approach. *SAR QSAR Environ Res* 1999; 10:117-129.
- (8) Maran U, Karelson M, Katritzky AR. A comprehensive QSAR treatment of the genotoxicity of heteroaromatic and aromatic amines. *Quant Struct -Act Relat* 1999; 18:3-10.
- (9) Basak SC, Mills D. Prediction of mutagenicity utilizing a hierarchical QSAR approach. *SAR QSAR Environ Res* 2001; 12:481-496.
- (10) Gramatica P, Consonni V, Pavan M. Prediction of aromatic amines mutagenicity from theoretical molecular descriptors. *SAR QSAR Environ Res* 2003; 14:237-250.
- (11) Mattioni BE, Kauffman GW, Jurs PC, Custer LL, Durham SK, Pearl GM. Predicting the genotoxicity of secondary and aromatic amines using data subsetting to generate a model ensemble. *J Chem Inf Comp Sci* 2003; 43:949-963.
- (12) Vracko M, Mills D, Basak SC. Structure-mutagenicity modelling using counter propagation neural networks. *Environ Toxicol Pharmacol* 2004; 16:25-36.
- (13) Cash GG. Prediction of the genotoxicity of aromatic and heteroaromatic amines using electrotopological state indices. *Mutat Res* 2001; 491:31-37.

- (14) Cash GG, Anderson B, Mayo K, Bogaczyk S, Tunkel J. Predicting genotoxicity of aromatic and heteroaromatic amines using electrotopological state indices. *Mutat Res* 2005; 585:170-183.
- (15) Lewis DFV, Ioannides C, Walker R, Parke DV. Quantitative structure-activity relationships and COMPACT analysis of a series of food mutagens. *Food Additives & Contaminants* 1995; 12:715-723.
- (16) Benigni R, Andreoli C, Giuliani A. QSAR models for both mutagenic potency and activity: application to nitroarenes and aromatic amines. *Environ Mol Mutagen* 1994; 24:208-219.
- (17) Hatch FT, Knize MG, Felton JS. Quantitative structure-activity relationships of heterocyclic amine mutagens formed during the cooking of food. *Environ Mol Mutagen* 1991; 17:4-19.
- (18) Hatch FT, Colvin ME, Seidl ET. Structural and quantum chemical factors affecting mutagenic potency of aminoimidazo-azaarenes. *Environ Mol Mutagen* 1996; 27:314-330.
- (19) Hatch FT, Colvin ME. Quantitative structure-activity (QSAR) relationships of mutagenic aromatic and heterocyclic amines. *Mutat Res* 1997; 376(1-2):87-96.
- (20) Felton JS, Knize MG, Hatch FT, Tanga MJ, Colvin ME. Heterocyclic amine formation and the impact of structure on their mutagenicity. *Cancer Letts* 1999; 143:127-134.
- (21) Hatch FT, Knize MG, Colvin ME. Extended quantitative structure-activity relationships for 80 aromatic and heterocyclic amines: structural, electronic, and hydrophathic factors affecting mutagenic potency. *Environ Mol Mutagen* 2001; 38:268-291.

1.1.2 Carcinogenic activity

- (22) Loew GH, Poulsen M, Kirkjian E, Ferrel J, Sudhindra BS, Rebagliati M. Computer-assisted mechanistic structure-activity: application to diverse classes of chemical carcinogens. *Environ Health Perspect* 1985; 61:69-96.
- (23) Benigni R, Giuliani A, Franke R, Gruska A. Quantitative structure-activity relationships of mutagenic and carcinogenic aromatic amines. *Chem Revs* 2000; 100:3697-3714.
- (24) Franke R, Gruska A, Giuliani A, Benigni R. Prediction of rodent carcinogenicity of aromatic amines: a quantitative structure-activity relationships model. *Carcinogenesis* 2001; 22:1561-1571.
- (25) Vracko M. A study of structure-carcinogenic potency relationship with artificial neural networks. The using of descriptors related to geometrical and electronic structures. *J Chem Inf Comput Sci* 1997; 37:1037-1043.
- (26) Gini G, Lorenzini M, Benfenati E, Grasso P, Bruschi M. Predictive carcinogenicity: a model for aromatic compounds, with nitrogen-containing substituents, based on molecular descriptors using an artificial neural network. *J Chem Inf Comput Sci* 1999; 39:1076-1080.

1.2 Aromatic Nitrocompounds

1.2.1 Nitroarenes

- (27) Biagi LG, Hrelia P, Gerra MG, Paolini M, Barbaro AM, Cantelli-Forti G. Structure-activity relationships of nitroimidazo (2,1-b) thiazoles in the salmonella mutagenicity assay. *Arch Toxicol* 1986; suppl.9:425-429.
- (28) Walsh DB, Claxton LD. Computer-assisted structure-activity relationships of nitrogenous cyclic compounds tested in Salmonella assays for mutagenicity. *Mutat Res* 1987; 182:55-64.
- (29) Maynard AT, Pedersen LG, Posner HS, McKinney JD. An Ab initio study of the relationship between nitroarene mutagenicity and electron affinity. *Mol Pharmacol* 1986; 29(6):629-636.
- (30) Compadre RL, Shusterman AJ, Hansch C. The role of hydrophobicity in the Ames test. The correlation of the mutagenicity of nitropolycyclic hydrocarbons with partition coefficients and molecular orbital indices. *Int J Quantum Chem* 1988; 34:91-101.
- (31) Lopez de Compadre RL, Debnath AK, Shusterman AJ, Hansch C. LUMO Energies and hydrophobicity as determinants of mutagenicity by nitroaromatic compounds in Salmonella typhimurium. *Environ Mol Mutagen* 1990; 15:44-55.
- (32) Debnath AK, de Compadre RLL, Debnath G, Shusterman A, Hansch C. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. Correlation with molecular orbital energies and hydrophobicity. *J Med Chem* 1991; 34:786-797.
- (33) Debnath AK, Lopez de Compadre RL, Shusterman AJ, Hansch C. Quantitative Structure-Activity Relationship investigation of the role of hydrophobicity in regulating mutagenicity in the Ames test: 2. Mutagenicity of aromatic and heteroaromatic nitro compounds in Salmonella typhimurium TA100. *Environ Mol Mutagen* 1992; 19:53-70.
- (34) King RD, Muggleton SH, Srinivasan A, Sternberg MJE. Structure-activity relationships derived by machine learning: the use of atoms and their bond connectivities to predict mutagenicity by inductive logic programming. *Proc Natl Acad Scie* 1996; 93(1):438-442.
- (35) Debnath AK, Hansch C. Structure-activity relationship of genotoxic polycyclic aromatic nitro compounds: further evidence for the importance of hydrophobicity and molecular orbital energies in genetic toxicity. *Environ Mol Mutagen* 1992; 20:140-144.
- (36) Debnath AK, Hansch C, Kim KH, Martin YC. Mechanistic interpretation of the genotoxicity of nitrofurans (antibacterial agents) using quantitative structure-activity relationships (QSAR) and comparative molecular field analysis (CoMFA). *J Med Chem* 1993; 36:1009-1116.
- (37) Caliendo G, Fattorusso C, Greco G, Novellino E, Perissutti E, Santagada V. Shape-dependent effects in a series of aromatic nitro compounds acting as mutagenic agents on *S. typhimurium* TA98. *SAR QSAR Environ Res* 1995; 4:21-27.
- (38) Fan M, Byrd C, Compadre CM, Compadre RL. Comparison of CoMFA models for Salmonella typhimurium TA98, TA100, TA98+S9 and TA100+S9 mutagenicity of nitroaromatics. *SAR QSAR Environ Res* 1998; 9:187.
- (39) Compadre RL, Byrd C, Compadre CM. Comparative QSAR and 3-D-QSAR analysis of the mutagenicity of nitroaromatic compounds. In: Devillers J, editor. *Comparative QSAR*. London: Taylor and Francis, Ltd, 1998: 111-136.

1.2.2 N-nitroso compounds

- (40) Singer GM, Andrews AW, Guo S. Quantitative structure-activity relationships of the mutagenicity of substituted n-nitroso-N- benzylmethyamines: possible implications for carcinogenicity. *J Med Chem* 1986; 29:40-44.
- (41) Hansch C, Leo A. QSAR of mutagenesis, carcinogenesis and antitumor drugs. 9-2-3. Nitrosoamines. *Exploring QSAR. Fundamentals and applications in chemistry and biology*. Washington, D.C.: American Chemical Society, 1995: 356-357.
- (42) Dunn III WJ, Wold S. An assessment of carcinogenicity of N-nitroso compounds by the SIMCA method of pattern recognition. *J Chem Inf Comput Sci* 1981; 21:8-13.
- (43) Frecer V, Miertus S. Theoretical QSAR study on carcinogenic potency of N-nitrosamines. *Neoplasma* 1988; 35:525-538.

1.3 Quinolines

- (44) Debnath AK, de Compadre RLL, Hansch C. Mutagenicity of quinolines in Salmonella typhimurium TA100, a QSAR study based on hydrophobicity and molecular orbital determinants. *Mutat Res* 1992; 280:55-65.
- (45) Smith CJ, Hansch C, Morton MJ. QSAR treatment of multiple toxicities: the mutagenicity and cytotoxicity of quinolines. *Mutat Res* 1997; 379:167-175.

1.4 Triazenes

- (46) Shusterman AJ, Debnath AK, Hansch C, Horn GW, Fronczek FR, Greene AC, Watkins SF. Mutagenicity of dimethyl heteroaromatic triazenes in the Ames test. The role of hydrophobicity and electronic effects. *Mol Pharmacol* 1989; 12:939-944.

1.5 Polycyclic aromatic hydrocarbons

- (47) Zhang L, Sannes K, Shusterman AJ, Hansch C. The structure-activity relationships of skin carcinogenicity of aromatic hydrocarbons and heterocycles. *Chem Biol Interact* 1992; 81:149-180.
- (48) Gallegos A, Robert D, Girones X, Carbo-Dorca R. Structure-toxicity relationships of polycyclic aromatic hydrocarbons using molecular quantum similarity. *J Comput -Aided Mol Design* 2001; 15:67-80.
- (49) Villemin D, Cherqaoui D, Mesbah A. Predicting carcinogenicity of polycyclic aromatic hydrocarbons from back-propagation neural network. *J Chem Inf Comput Sci* 1994; 34:1288-1293.
- (50) Richard AM, Woo YT. A CASE-SAR analysis of polycyclic aromatic hydrocarbon carcinogenicity. *Mutat Res* 1990; 242:285-303.

1.6 Halogenated aliphatics

- (51) Benigni R, Andreoli C, Conti L, Tafani P, Cotta-Ramusino M, Carere A, Crebelli R. Quantitative structure-activity relationship models correctly predict the toxic and aneuploidizing properties of six halogenated methanes in *Aspergillus nidulans*. *Mutag* 1993; 8:301-305.
- (52) Crebelli R, Andreoli C, Carere A, Conti L, Crochi B, Cotta-Ramusino M, Benigni R. Toxicology of halogenated aliphatic hydrocarbons: structural and molecular determinants for the disturbance of chromosome segregation induction of lipid peroxidation. *Chem Biol Interact* 1995; 98:113-129.
- (53) Basak SC, Balasubramanian K, Gute BD, Mills D, Gorcynska A, Roszak S. Prediction of cellular toxicity of halocarbons from computed chemodescriptors: a hierarchical QSAR approach. *J Chem Inf Comput Sci* 2003.

1.7 Direct acting compounds

1.7.1 Platinum amines

- (54) Hansch C, Venger BH, Panthanickal A. Mutagenicity of substituted (o-phenyldiamine)platinum dichloride in the Ames test. A quantitative structure-activity analysis. *J Med Chem* 1980; 23:459.

1.7.2 Furanones

- (55) LaLonde RT, Leo H, Perakyla H, Dence CW, Farrell RP. Associations of the bacterial mutagenicity of halogenated 2(5H)-furanones with their MNDO-PM3 computed properties and mode of reactivity with sodium borohydride. *Chem Res Toxicol* 1992; 5:392-400.
- (56) Tuppurainen K. Frontier orbital energies, hydrophobicity and steric factors as physical QSAR descriptors of molecular mutagenicity. A review with a case study: MX compounds. *Chemosphere* 1999; 38:3015-3030.

1.7.3 Epoxides

- (57) Hooberman BH, Chakraborty PK, Sinsheimer JE. Quantitative structure-activity relationships for the mutagenicity of propylene oxides with *Salmonella*. *Mutat Res* 1993; 299:85-93.
- (58) Sugiura K, Goto M. Mutagenicity of styrene oxide derivatives on bacterial test systems: relationship between mutagenic potency and chemical reactivity. *Chem Biol Interact* 1981; 35:71-91.

- (59) Tamura N, Takahashi K, Shirai N, Kawazoe Y. Studies on chemical carcinogens XXI. Quantitative structure-mutagenicity relationships among substituted styrene oxides. *Chem Pharm Bull* 1982; 30:1393-1400.

1.8 Aliphatic aldehydes

- (60) Benigni R, Passerini L, Rodomonte A. Structure-activity relationships for the mutagenicity and carcinogenicity of simple and α - β unsaturated aldehydes. *Environ Mol Mutagen* 2003; 42:136-143.
- (61) Benigni R, Conti L, Crebelli R, Rodomonte A, Vari' MR. Simple and α - β -unsaturated aldehydes: correct prediction of genotoxic activity through Structure-Activity Relationship models. *Environ Mol Mutagen* 2005; 46:268-280.

2. (Q)SARs for noncongeneric sets of chemicals

2.1 Mutagenicity

- (62) Ashby J. Fundamental structural alerts to potential carcinogenicity or noncarcinogenicity. *Environ Mutagen* 1985; 7:919-921.
- (63) Kazius J, McGuire R, Bursi R. Derivation and Validation of Toxicophores for Mutagenicity Prediction. *J Med Chem* 2005; 48:312-320.
- (64) Kazius J, Nijssen S, Kok J, Back T, Ijzerman AP. Substructure mining using elaborate chemical representation. *J Chem Inf Model* 2006; in press.
- (65) Bailey AB, Chanderbhan N, Collazo-Braier N, Cheeseman MA, Twaroski ML. The use of structure-activity relationship analysis in the food contact notification program. *Regulat Pharmacol Toxicol* 2005; 42:225-235.
- (66) Helma C, Cramer T, Kramer S, De Raedt L. Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of noncongeneric compounds. *J Chem Inf Comp Sci* 2004; 44:1402-1411.
- (67) Lewis DFV, Ioannides C, Parke DV. A quantitative structure-activity relationship (QSAR) study of mutagenicity in several series of organic chemicals likely to be activated by Cytochrome P450 enzymes. *Teratog Carcinog Mutagen* 2003; 1:187-193.
- (68) Basak SC, Mills D, Gute BD, Hawkins DM. Predicting mutagenicity of congeneric and diverse sets of chemicals using computed molecular descriptors: a hierarchical approach. In: Benigni R, editor. *Quantitative Structure-Activity Relationship (QSAR) models of chemical mutagens and carcinogens*. Boca Raton: CRC Press, 2003: 207-234.
- (69) Brinn M, Walsh P, Payne M, Bott B. Neural network classification of mutagens using structural fragment data. *SAR QSAR Environ Res* 1992; 1:169-211.

- (70) Contrera JF, Matthews EJ, Kruhlak NL, Benz RD. In silico screening of chemicals for bacterial mutagenicity using electropological E-state indices and MDL QSAR software. *Regulat Pharmacol Toxicol* 2005; 43:313-323.
- (71) Maran U, Sild S. QSAR modeling of genotoxicity on non-congeneric sets of organic compounds. *Artific Intell Rev* 2003; 20:13-38.
- (72) Mekenyan O, Dimitrov S, Serafimova R, Thompson ED, Kotov S, Dimitrova N, Walker JD. Identification of the structural requirements for mutagenicity by incorporating molecular flexibility and metabolic activation of chemicals I: TA100 model. *Chem Res Toxicol* 2004; 17:753-766.
- (73) Votano JR, Parham M, Hall LH, Kier LB, Oloff S, Tropsha A, Xie Q, Tong W. Three new consensus QSAR models for the prediction of Ames genotoxicity. *Mutag* 2005; 19:365-377.

2.2 Carcinogenicity

- (74) Helma C. Lazy Structure-Activity Relationships (lazar) for the prediction of Rodent Carcinogenicity and Salmonella mutagenicity. *Molecular Diversity*, 2006; in press.
<http://www.predictive-toxicology.org/lazar/index.html>
- (75) Matthews EJ, Contrera JF. A new highly specific method for predicting the carcinogenic potential of pharmaceuticals in rodents using enhanced MCASE QSAR-ES software. *Regulat Pharmacol Toxicol* 1998; 28:242-264.
- (76) Contrera JF, Matthews EJ, Benz RD. Predicting the carcinogenic potential of pharmaceuticals in rodents using molecular structural similarity and E-state indices. *Regulat Toxicol Pharmacol* 2003; 38:243-259.
- (77) King RD, Srinivasan A. Prediction of Rodent Carcinogenicity Biossays from Molecular Structure Using Inductive Logic programming. *Environ Health Perspect* 1996; 104(suppl.5):1031-1040.
- (78) Lewis DFV, Ioannides C, Parke DV. Validation of a novel molecular orbital approach COMPACT for the prospective safety evaluation of chemicals, by comparison with rodent carcinogenicity and Salmonella mutagenicity data evaluated by the U.S. NCI/NTP. *Mutat Res* 1993; 291:61-77.

8 Appendix 2 Scoring results for selected models

		Unambiguous algorithm					Applicability	Statistics			Mechanism	Others	
	Model	A1	A2.1	A2.2	A2.3	A3	B	C1	C2	C3	D	E1	E2
1.1.	Aromatic Amines												115
1.1.1													
	4	1	1	1	1	1	2	1	1 (ref. 5)	2	1	95	
	6	1	1	1	2	4	3	1	1	2	2	73	
	7	1	1	1	2	4	3	1	1	2	2	127	
	8	2	1	1	2	3	3	1	1	2	2	95	
	9	1	1	1	2	4	3	1	1	2	2	95	
	10	1	1	1	2	2	3	1	1	2	2	95	
	11	2	1	1	2	2	3	1	1	1	2	334	
	12	2	1	1	2	4	3	1	1	2	2	95	
	13	1	1	1	2	2	3	1	1	1 (ref. 14)	2	95	
	21	1	1	1	1	2	2	1	2	2	1	80	
1.1.2													
	23	1	1	1	1	1	2	1	2	2	1	58	
	24	1	1	1	1	1	2	1	1 (ref. 5)	2	1	82	
	25	2	1	1	2	4	3	1	1	2	2	45	
	26	2	1	1	2	2	3	1	1	2	2	104	
1.2.1	Nitroarenes												42
	28	2	1	1	2	3	3	1	1	2	2	114	
	32	1	1	1	1	1	3	1	2	2	1	197	
	33	1	1	1	1	1	2	1	2	2	1	132	
	34	1	1	1	1	3	3	1	1	2	2	188+42	
	35	1	1	1	2	1	2	1	2	2	1	23	
	36	1	1	1	2	2	2	1	2	2	1	46	
1.2.2	N-nitroso												0
	42	1	1	1	2	1	2	1	2	2	2	61	
1.5.	PAH												6
	47	1	1	1	1	1	2	1	2	2	1	239	
	48	2	1	1	2	4	3	1	1	2	2	78	
	49	2	1	1	2	4	3	1	1	2	2	94	
	50	1	1	1	1	3	2	1	1	1	1	102	

1.6.	Halogenated Aliphatics												113
	52	1	1	1	1	2	3	1	2	1	1	55	
	53	2	1	1	2	4	3	1	1	2	2	55	
1.7.2	Furanones												3
	56	1	1	1	1	1	2	1	2	2	1	24	
1.8.	Aliphatic Haldehydes												33
	60	1	1	1	1	1	3	1	1	1 (ref. 61)	1	29	
2.1.	Non congeneric mutagens												
	62	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	1	N.A.	
	63	1	1	1	1	1	3	1	2	1	1	4337	
	64	1	1	1	1	2	3	1	1	2	1	4069	
	65	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	1	N.A.	
	68	1	2	2	2	4	3	1	1	2	2	508	
	69	2	2	2	2	4	3	1	1	2	2	607	
	70	2	2	2	2	3	2	1	1	2	2	3338	
	71	1	1	1	2	3	3	1	1	2	2	177+212	
	72	1	2	2	2	3	1	1	1	2	2	336	
	73	2	2	2	2	3	2	1	1	2	2	3363	
2.2.	Non congeneric carcinogens												
	74	1	1	1	2	3	1	1	1	1	2	1447	
	76	2	2	2	2	3	2	1	1	1	2	1275	
	77	1	1	1	1	3	2	1	1	2	2	330	

9 Appendix 3 Regression-based models for mutagenicity and carcinogenicity

Ref. 4 (Debnath et al., 1992)

Mutagenicity of aromatic amines in *S. typhimurium* TA98 and TA100 strains, with S9 metabolic activation.

$$\log\text{TA98} = 1.08 (\pm 0.26) \log P + 1.28 (\pm 0.64) \text{HOMO} - 0.73 (\pm 0.41) \text{LUMO} + 1.46 (\pm 0.56) \text{I}_L + 7.20 (\pm 5.4)$$

$$n=88 \quad r=0.898 \quad s=0.860$$

$$\log\text{TA100} = 0.92 (\pm 0.23) \log P + 1.17 (\pm 0.83) \text{HOMO} - 1.18 (\pm 0.44) \text{LUMO} + 7.35 (\pm 6.9)$$

$$n=67 \quad r=0.877 \quad s=0.708$$

$\log\text{TA98}$ and $\log\text{TA100}$: mutagenic potency as $\log(\text{revertants/nmol})$

$\log P$: logarithm of the octanol / water partition coefficient

HOMO: energy of the Highest Occupied Molecular Orbital

LUMO: energy of the Lowest Unoccupied Molecular Orbital

I_L : indicator variable, 1 for compounds with three or more fused rings

Ref 10 (Gramatica et al., 2003)

Mutagenicity of aromatic amines in *S. typhimurium* TA98 and TA100 strains, with S9 metabolic activation.

$$\log\text{TA98} = -3.98 + 2.40 \text{MWC07} + 0.56 \text{MATS7m} + 2.44 \text{Mor27u} + 1.12 \text{Mor15m}$$

$$n = 60; r^2 = 80.3; Q^2_{\text{LOO}} = 76.6; Q^2_{\text{LMO}} = 75.9; Q^2_{\text{ext}} = 68.9; K_{\text{xx}} = 27.9;$$

$$s = 0.827; F_{(55)} = 55.87; \text{SDEC} = 0.791; \text{SDEP} = 0.861; \text{SDEP}_{\text{ext}} = 0.991$$

$$\log TA_{100} = -3.99 - 0.61 \text{ nHA} + 9.55 \text{ ATS5p} + 0.65 \text{ L2v}$$

$$n = 46; r^2 = 81.2; Q^2_{\text{LOO}} = 78.0; Q^2_{\text{LMO}} = 77.4; Q^2_{\text{ext}} = 67.1; K_{\text{xx}} = 17.1;$$

$$s = 0.579; F_{(42)} = 60.40; \text{SDEC} = 0.553; \text{SDEP} = 0.598; \text{SDEP}_{\text{ext}} = 0.731$$

The definitions of the parameters are in the original references quoted in Ref. 10.

Ref 23 (Benigni et al., 2000)

Carcinogenic potency of aromatic amines in rodents

$$\text{BRM} = 0.88(\pm 0.27) \log P * I(\text{monoNH}_2) + 0.29(\pm 0.20) \log P * I(\text{diNH}_2) + 1.38(\pm 0.76) \text{HOMO} - 1.28(\pm 0.54) \text{LUMO} - 1.06(\pm 0.34) \text{EMR}_{2,6} - 1.10(\pm 0.80) \text{MR}_3 - 0.20(\pm 0.16) \text{Es(R)} + 0.75(\pm 0.75) I(\text{diNH}_2) + 11.16(\pm 6.68)$$

$$n = 37 \quad r = 0.907 \quad r^2 = 0.823 \quad s = 0.381 \quad F = 16.3 \quad P < 0.001$$

$$\text{BRR} = 0.35(\pm 0.18) \log P + 1.93(\pm 0.48) I(\text{Bi}) + 1.15(\pm 0.60) I(\text{F}) - 1.06(\pm 0.53) I(\text{BiBr}) + 2.75(\pm 0.64) I(\text{RNNO}) - 0.48(\pm 0.30)$$

$$n = 41 \quad r = 0.933 \quad r^2 = 0.871 \quad s = 0.398 \quad F = 47.4 \quad P < 0.001$$

$$\text{BRM} = \log (\text{MW}/\text{TD}_{50})_{\text{mouse}}$$

$$\text{BRR} = \log (\text{MW}/\text{TD}_{50})_{\text{rat.}}$$

TD50: daily dose required to halve the probability for an experimental animal of remaining tumorless to the end of its standard life span.

EMR_{2,6}: sum of Molar Refractivity of substituents in the *ortho*-positions of the aniline ring;

MR₃, Molar Refractivity of substituents in the meta-position of the aniline ring;

Es(R), Charton's substituent constant for substituents at the functional amino group;

I(monoNH₂) = 1 for compounds with only one amino group;

$I(\text{diNH}_2) = 1$ for compounds with more than one amino group;
 $I(\text{Bi}) = 1$ for biphenyls;
 $I(I(\text{BiBr})) = 1$ for biphenyls with a bridge between the phenyl rings;
 $I(\text{RNNO}) = 1$ for compounds with the group $\text{N}(\text{Me})\text{NO}$;
 $I(\text{F}) = 1$ for aminofluorenes.

Ref 24 (Franke et al., 2001)

Carcinogenicity of aromatic amines in rodents (discriminant analysis)

$$w = -2.86 L(R) + 2.65 B5(R) - 1.16 \text{HOMO} + 1.76 \text{LUMO} + 0.40 \text{MR3} + 0.58 \text{MR5} + 0.54 \text{MR6} - 1.55 I(\text{An}) + 0.74 I(\text{NO}_2) - 0.55 I(\text{BiBr})$$

$$w(\text{mean}, \text{Class1}) = -1.56 \quad N1 = 13$$

$$w(\text{mean}, \text{Class2}) = 0.38 \quad N2 = 53$$

$N1$ = number of non-carcinogens (Class 1);

$N2$ = number of carcinogens (Class 2);

$L(R)$: Sterimol length;

$B5(R)$ Sterimol maximal width;

MR3 , MR5 , MR6 : MR contributions of substituents in position 3, 5, and 6 to the amino group;

$I(\text{An})$: 1 for anilines;

$I(\text{NO}_2)$: 1 for the presence of a NO_2 group;

$I(\text{BiBr})$: 1 for biphenyls with a bridge between the phenyl rings;

Ref 32 (Debnath et al., 1991)

Mutagenicity of aromatic and heteroaromatic nitro compounds in *S. typhimurium* strain TA98

$$\log \text{TA98} = 0.65(\pm 0.16) \log P - 2.90(\pm 0.59) \log (\beta 10^{\log P} + 1) - 1.38(\pm 0.25) \text{LUMO} + 1.88(\pm 0.39) I_1 - 2.89(\pm 0.81) I_a - 4.15(\pm 0.58)$$

$$n=188, r=0.900, s=0.886, \log P_0=4.93, \log \beta=5.48, F_{1,181}=48.6$$

I_1 : 1 for compounds with 3 or more fused rings;

I_a : 1 for 5 substances of the set that are much less active than expected.

Ref 33 (Debnath et al., 1992b)

Mutagenicity of nitroarenes in *S. typhimurium* TA100, without metabolic activation

$$\begin{aligned}\log \text{TA100} = & 1.20(\pm 0.15)\log P - 3.40(\pm 0.74)\log(\beta 10^{\log P} + 1) - 2.05(\pm 0.32)\text{LUMO} - \\ & 3.50(\pm 0.82) I_a + 1.86(\pm 0.74)I_{\text{ind}} - 6.39(\pm 0.73) \\ n = & 117, r = 0.886, s = 0.835, \log P_0 = 5.44(\pm 0.24), \log \beta = -5.7, F_{1,110} = 24.7\end{aligned}$$

I_a : 1 for compounds where acenthrylene ring is present;

I_{ind} : 1 for the 1- and 2-methylindazole derivatives.

Ref 35 (Debnath and Hansch, 1992)

Mutagenicity of polycyclic aromatic nitro compounds in the SOS chromotest in *Escherichia coli* PQ37.

$$\begin{aligned}\log \text{SOSIP} = & 1.07 (\pm 0.36) \log P - 1.57 (\pm 0.57) \text{LUMO} - 6.41 (\pm 1.8) \\ n = & 15, r = 0.922, s = 0.534, F_{1,12} = 36.21\end{aligned}$$

SOSIP: SOS induction factor/ nmole.

Ref 36 (Debnath et al., 1993)

Genotoxicity of nitrofurans in the SOS chromotest in *Escherichia coli* PQ37.

$$\begin{aligned}\log \text{SOSIP} = & -33.1(\pm 11.9) q_{c2} + 1.00(\pm 0.26)\log P - 1.50(\pm 0.49) I_{\text{sat}} - 1.19(\pm 0.49)\text{MR} - \\ & 0.76(\pm 0.49)I_{5,6} - 3.76(\pm 1.56) \\ n = & 40, r = 0.900, s = 0.475, F_{1,34} = 9.76\end{aligned}$$

q_{c2} : partial atomic charge on the carbon attached to the nitro group;

I_{sat} : 1 for saturated ring compounds;

$I_{5,6}$: 1 for compounds with substituents at the 5- or 6- position of 2-nitronaphthofurans and pyrenofurans.

Ref 47 (Zhang et al., 1992)

Skin carcinogenicity of polycyclic aromatic hydrocarbons

$$\log I_{ball} = 0.55(\pm 0.09)\log P - 1.17(\pm 0.14) \log (\beta 10^{\log P} + 1) + 0.39(\pm 0.11)LK + 0.47(\pm 0.26)HOMO + 1.93(\pm 2.4)$$

$n=161, r=0.845, s=0.350, \log P_0 = 6.67(\pm 0.217), \log \beta = -6.81, F_{1,155}=12.8$

I_{ball} index = (Tumor incidence) (100%) / mean latent period in days,

where Tumor incidence = number of animal with tumors / number of animals alive when the first tumor appears.

LK: 1 for compounds with a substituent attached to a L or K region.

Ref 52 (Crebelli et al., 1995)

Induction of aneuploidy by halogenated aliphatics in *A. nidulans*

$$\log (1/LEC) = 0.83 + 0.07 MR - 4.91 LUMO - 3.41 DIFF$$

$$N = 24; F_{tot} = 69.07$$

LEC: Lowest Effective Concentration;

DIFF: LUMO – HOMO.

Ref 56 (Tuppurainen, 1999)

Mutagenicity of halogenated furanones (lactones) in *S. typhimurium* TA100

$$\ln \text{TA100} = -12.7(\pm 1.1) \text{LUMO} - 12.0(\pm 1.3) \\ n=24, r=0.930, s=1.33, F=141.0$$

Ref 60 (Benigni et al., 2003)

Mutagenicity of α - β unsaturated aldehydes in *S. typhimurium* TA100

$$\log \text{TA100} = -12.61592 - 4.58430 \text{ LUMO} - 3.66205 \text{ MR} + 72.46140 \text{ C-carb} + 2.55239 \log P + \\ 13.09442 \text{ C-}\beta \\ n = 17; r^2 = 0.84; q^2 = 0.40$$

C-carb= partial charge on the carbonilic carbon;

C- β = partial charge on the β carbon.

Mutagenicity of α - β unsaturated aldehydes in *S. typhimurium* TA100 (discriminant analysis)

$$\text{Negatives} = -47.13331 + 38.24641 \text{ MR} - 31.77763 \log P + 30.46799 \text{ LUMO} \\ \text{Positives} = -20.52153 + 25.41469 \text{ MR} - 21.45102 \log P + 19.77513 \text{ LUMO}$$

$n = 20$; 100% correct reclassification; 3/20 errors in cross-validation.

European Commission

22772 EN – DG Joint Research Centre, Institute IHCP

Title: Collection and Evaluation of (Q)SAR Models for Mutagenicity and Carcinogenicity

Authors: Romualdo Benigni, Cecilia Bossa, Tatiana Netzeva, Andrew Worth

Luxembourg: Office for Official Publications of the European Communities

2007 – 119 pp. – 21 x 29.7 cm

EUR - Scientific and Technical Research series; **ISSN 1018-5593**

Abstract

This evaluation of the non-commercial (Q)SARs for mutagenicity and carcinogenicity consisted of a preliminary survey (Phase I), and then of a more detailed analysis of short listed models (Phase II). In Phase I, the models were collected from the literature, and then assessed according to the OECD principles –based on the information provided by the authors-. Phase I provided the support for short listing a number of promising models, that were analyzed more in depth in Phase II. In Phase II, the information provided by the authors was completed and complemented with a series of analyses aimed at generating an overall profile of each of the short listed models.

The models can be divided into two families based on their target: a) congeneric; and b) non-congeneric sets of chemicals.

The QSARs for congeneric chemicals include most of the chemical classes top ranking in the EU High Production Volume list, with the notable exception of the halogenated aliphatics. They almost exclusively aim at modeling *Salmonella* mutagenicity and rodent carcinogenicity, which are crucial toxicological endpoints in the regulatory context. The lack of models for *in vivo* genotoxicity should be remarked. Overall the short listed models can be interpreted mechanistically, and agree with, and/or support the available scientific knowledge, and most of the models have good statistics. Based on external prediction tests, the QSARs for the potency of congeneric chemicals are 30 to 70 % correct, whereas the models for discriminating between active and inactive chemicals have considerably higher accuracy (63 to 100 %), thus indicating that predicting intervals is more reliable than predicting individual data points. The internal validation procedures (e.g., cross-validation, etc...) did not seem to be a reliable measure of external predictivity.

Among the non-local, or global approaches for non-congeneric data sets, four models based on the use of Structural Alerts (SA) were short listed and investigated in more depth. The four sets did not differ to a large extent in their performance. In the “general” databases of chemicals the SAs appear to agree around 65% with rodent carcinogenicity data, and 75% with *Salmonella* mutagenicity data. The SAs based models do not seem to work equally efficiently in the discrimination between active and inactive chemicals within individual chemical classes. Thus, their main role is that of preliminary, or large-scale screenings. A priority for future research on the SAs is their expansion to include alerts for nongenotoxic carcinogens.

A general indication of this study, valid for both congeneric and noncongeneric models, is that there is uncertainty associated with (Q)SARs; the level of uncertainty has to be considered when using (Q)SAR in a regulatory context. However, (Q)SARs are not meant to be black-box machines for predictions, but have a much larger scope including organization

and rationalization of data, contribution to highlight mechanisms of action, complementation of other data from different sources (e.g., experiments). Using only non-testing methods, the larger the evidence from QSARs (several different models, if available) and other approaches (e.g. chemical categories, read across) the higher the confidence in the prediction.

The mission of the JRC is to provide customer-driven scientific and technical support for the conception, development, implementation and monitoring of EU policies. As a service of the European Commission, the JRC functions as a reference centre of science and technology for the Union. Close to the policy-making process, it serves the common interest of the Member States, while being independent of special interests, whether private or national.